Protein-Protein Interaction Networks and Impact of Disease-related Mutations.

Lu, Grace

Awarding institution:
King's College London

KING'S COLLEGE LONDON

Protein-Protein Interaction Networks and Impact of Disease-related Mutations.

by

Hui-Chun Lu

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
School of Biomedical and Health Sciences
Randall Division of Cell and Molecular Biophysics

January 2014

Declaration of Authorship

I, Hui-Chun Lu, declare that this thesis titled, 'Protein-Protein Interaction Networks and Impact of Disease-related Mutations' and the work presented in it are my own, I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main source of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

KING'S COLLEGE LONDON

Abstract

School of Biomedical and Health Sciences
Randall Division of Cell and Molecular Biophysics

Doctor of Philosophy

by Hui-Chun Lu

Numerous studies have suggested the correlation between the stability of protein complexes, the importance of Protein Protein Interactions (PPIs) and the resulting molecular mechanisms for the underlying protein functions. In particular, the three-dimensional (3D) properties of protein binding interfaces are thought to embed key roles in mediating biological activities and in regulating cellular functions. The alteration of binding interfaces can disrupt the biological system in cell and consequently result in different phenotypic traits or diseases. In a scenario in which the rapid growth of biologically relevant information contributed by large-scale sequencing projects has paved the way to insights into the relationship between genotype and phenotype, it is really important to effectively combine all available information playing a role in this, to extract some principle rules guiding our understanding on the occurring molecular and atomistic mechanisms.

It is therefore timely to implement large-scale studies on the role of gene variants on Protein-Protein Interaction Networks (PPINs) and more specifically protein complexes. The objective of this project is a proteome-wide scale analysis of Protein Interaction data by mapping human genetic variation data onto structurally determined binary protein complexes. 3D PPINs were used as a tool to extract the association between human proteins and to enable an insight into molecular features of human genetic variation. A

comprehensive literature review on the topic of PPINs is given in Chapter 1 ("Introduction to Protein-Protein Interactions and Networks"). Non-synonymous Single Nucleotide Polymorphisms (nsSNPs) were the main focus in this study since they can directly cause conformational changes of proteins or failures in forming protein complexes. Two disease-related nsSNP datasets were investigated including: a) germ-line disease nsSNPs and b) somatic cancer nsSNPs. A set of nsSNPs which are known not to be related to diseases was used as the background to be compared with the chosen disease nsSNPs in order to highlight the characteristic properties identifying the features of disease nsSNPs. A survey on human genetic variation and the current state of related studies is presented in Chapter 2 ("Human Gene Variants"). The study of inter-domain disordered regions was also included in this project, as recent studies suggested their importance in regulating biological functions. An introduction on protein "Intrinsic Disorder" is also presented in Chapter 2.

An automated system pipeline was developed in this study to generate structure-integrated PPINs at protein domain level, map nsSNPs onto these structures, and classify nsSNPs. The collected nsSNP datasets are classified by their occurrence in different protein regions, including surface, interface, core and disordered. The detail of the pipeline development is given in Chapter 4 ("Pipeline to Generate 3D Protein-Protein Interaction Networks"). The interface regions showed a previously documented enrichment with disease-related nsSNPs. In addition, our results showed that germ-line disease nsSNPs and somatic cancer nsSNPs exhibit distinctive features in terms of their physical-chemical preferences and functional specificity. This may suggest that these two types of disease-related nsSNPs affect cellular functions through different mechanisms. Moreover, the functions of affected proteins were found to be highly related to the types of diseases the germ-line nsSNPs lead to. These

results will be presented together in Chapter 3 ("Computational Analyses of Disease-related Variants").

Published Articles

1. Setta-Kaffetzi, N., Simpson M. A., Navarini A. A., Patel V. M., Lu, H.-C., Allen M. H., Duckworth M., Bachelez H., Burden A. D., Choon A.-E., Griffiths C., Kirby B., Kolios A., Seyger M., Prins C., Smahi A., Trembath R., Fraternali F., Smith C., Barker J. N. & Capon F. AP1S3 mutations are associated with pustular psoriasis and impaired Toll-like receptor 3 trafficking. *The American Journal of Human Genetics* 94, 790-797 (2014).

2. Scharner, J., Lu, H.-C., Fraternali, F., Ellis, J. A. & Zammit, P. S. Mapping disease-related missense mutations in the immunoglobulin-like fold domain of lamin A/C reveals novel genotype-phenotype associations for laminopathies. *Proteins* (2013). doi:10.1002/prot.24465

3. Lu, H.-C., Fornili, A. & Fraternali, F. Protein-protein interaction networks studies and importance of 3D structure knowledge. *Expert Review of Proteomics* 10, 511-520 (2013).

4. Fornili, A., Pandini, A., Lu, H.-C. & Fraternali, F. Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *Journal of Chemical Theory and Computation* 9, 5127-5147 (2013).

5. Baines, A. J., Lu, H.-C. & Bennett, P. M. The Protein 4.1 family: Hub proteins in animals for organizing membrane proteins. *Biochimica et Biophysica Acta* 1838(2), 605-619 (2013).

6. Satoh, T., Smith, A., Sarde, A., Lu, H.-C., Mian, S., Mian, S., Trouillet, C., Mufti, G., Emile, J.-F., Fraternali, F., Donadieu, J. & Frederic Geissmann. B-RAF mutant alleles associated with Langerhans cell histiocytosis, a granulomatous pediatric

disease. *PLOS ONE* 7, e33891 (2012).

Acknowledgements

I would first thank Prof. Franca Fraternali for her supervision and guidance. Her support and encouragement had helped me get through these few years of the PhD project. I am also very thankful for her kind caring in every way. It had been a very pleasant experience working in such a group with nice atmosphere.

I would also like to thank Dr. Arianna Fornili whose kind help was utterly precious in understanding biological terms, as well as in the data analysis. The discussions with her were always very interesting and fruitful. I am also very thankful to Arianna and Dr. Jens Kleinjung for their help in the redaction of this thesis and giving advice to improve my work.

I would like to thank everyone in the Fraternali group and the ex-members of the group for the help and their friendship. (Flavia Autore, Alessandro Padini, Aisling Williams, Luis Fernandes, Pierre Martinez, Nesrine Chakroun, Pietro Buffa) Also many many thanks to many people in the Randall Division for their friendship, constant support and help in many ways, in particular, Virginia Tajadura, Gian Felice, Marie Pang, Celine Wu, Dr. Balvinder Dhaliwal, Helen Rudkin and my afternoon coffee-break friends (Daniel O'Loughlin, Flavia Autore, Norhakim Yahya).

Finally, I want to thank my family and my friends for their support. I especially thank my parents for their fully support and understanding for being unable to show up in many family occasions.

Contents

List of Figures

List of Tables

List of Abbreviations

3D Three-Dimensional

CNVs Copy Number Variations

GD Germ-line Disease

GO Gene Ontology

GWAS Genome-Wide Association Studies

HMMs Hidden Markov Models

IDRs Intrinsically Disordered Regions

LCAN Lowest Common Ancestor Node

LCH Langerhans Cell Histiocytosis

MAF Minor Allele Frequency

MoRFs Molecular Recognition Features

nsSNPs non-synonymous Single-Nucleotide Polymorphisms

PPIs Protein-Protein Interactions

PPINs Protein-Protein Interaction Networks

PTM Post-Translational Modification

SASA Solvent Accessible Surface Area

SC Somatic Cancer

SNPs Single-Nucleotide Polymorphisms

SNVs Single Nucleotide Variations

Chapter 1

Introduction to Protein-Protein Interactions and Networks

Protein-Protein Interaction Networks (PPINs) are a powerful tool to study biological processes in living cells. In this chapter, we present the progress of PPIN studies from abstract to more detailed representations. We will focus on 3D interactome networks, which offer detailed information at the atomic level. This information can be exploited in understanding not only the underlying cellular mechanisms, but also how human variants and disease-causing mutations affect protein functions and complexes' stability. Recent studies have used structural information on PPINs to also understand the molecular mechanisms of binding partner selection. We will address the challenges in generating 3D PPINs due to the restricted number of solved protein structures. Finally, some of the current use of 3D PPINs will be discussed, highlighting their contribution to the studies in genotype–phenotype relationships and in the optimization of targeted studies to design novel chemical compounds for medical treatments.

This chapter has been published as a review paper [1] which is attached in Appendix B. The first draft of the review was completed by the first author under supervision of Prof. Franca Fraternali. The content contributed by other authors to the published review includes:

- Mapping of functional pathway for cancer studies (Page 5 paragraph 1 of this thesis).
- Protein allosteric regulation (Page 6 from line 22 to the end of the paragraph, and Page 7 line 15 of this thesis).

1.1 Introduction

Most biological processes in a living cell, such as transcription regulation, signal transduction and cell motility, are mediated by Protein–Protein Interactions (PPIs). Network representations effectively address the complexity of PPIs in biological systems. Indeed, networks provide a highly compact and comprehensive view of binary relationships, in which nodes represent proteins/genes and edges indicate functional association or physical interactions between protein/gene pairs (Figure 1a). Many studies used network analysis to report proteins/genes involved in a particular disease or function [2, 3], and to complement large–scale siRNA screenings [4, 5]. The network topological properties of genes, including the degree of interactions, the clustering coefficient and the betweenness, which are the measurements of connectivity, interconnectivity and centrality, respectively, are often used to characterize topological features of a network [6, 7] and provide useful insights: for example, cancer proteins were found to have distinguishable topological features to the other proteins. These are generally found to be hub proteins with relatively high centrality [8, 9]. Network representations can also be used to study the associations between objects.

In the study of Goh *et al.* [10], a human diseasome bipartite network was constructed to model associations between genes and diseases. A link between a disease and a disease gene indicates that mutations in that gene are resulting in the specific disease. Many diseases were found to share a common genetic origin. This was an interesting finding which suggested that diseases may not be as independent of each other as we know from the traditional clinical assessment.

Apart from disease-related studies, PPINs are also often used in functional studies to assign putative functions to newly discovered genes [11, 12] using algorithms based on the Guilt-by-Association principle (Figure 1B). This has been particularly explored for plant-specific proteins [13, 14, 15] as the majority of these protein functions remain unknown, and yet it is critical to understand their biological relevance and the involving biological processes, including growth control and genotype–phenotype relationships for the variety of plants.

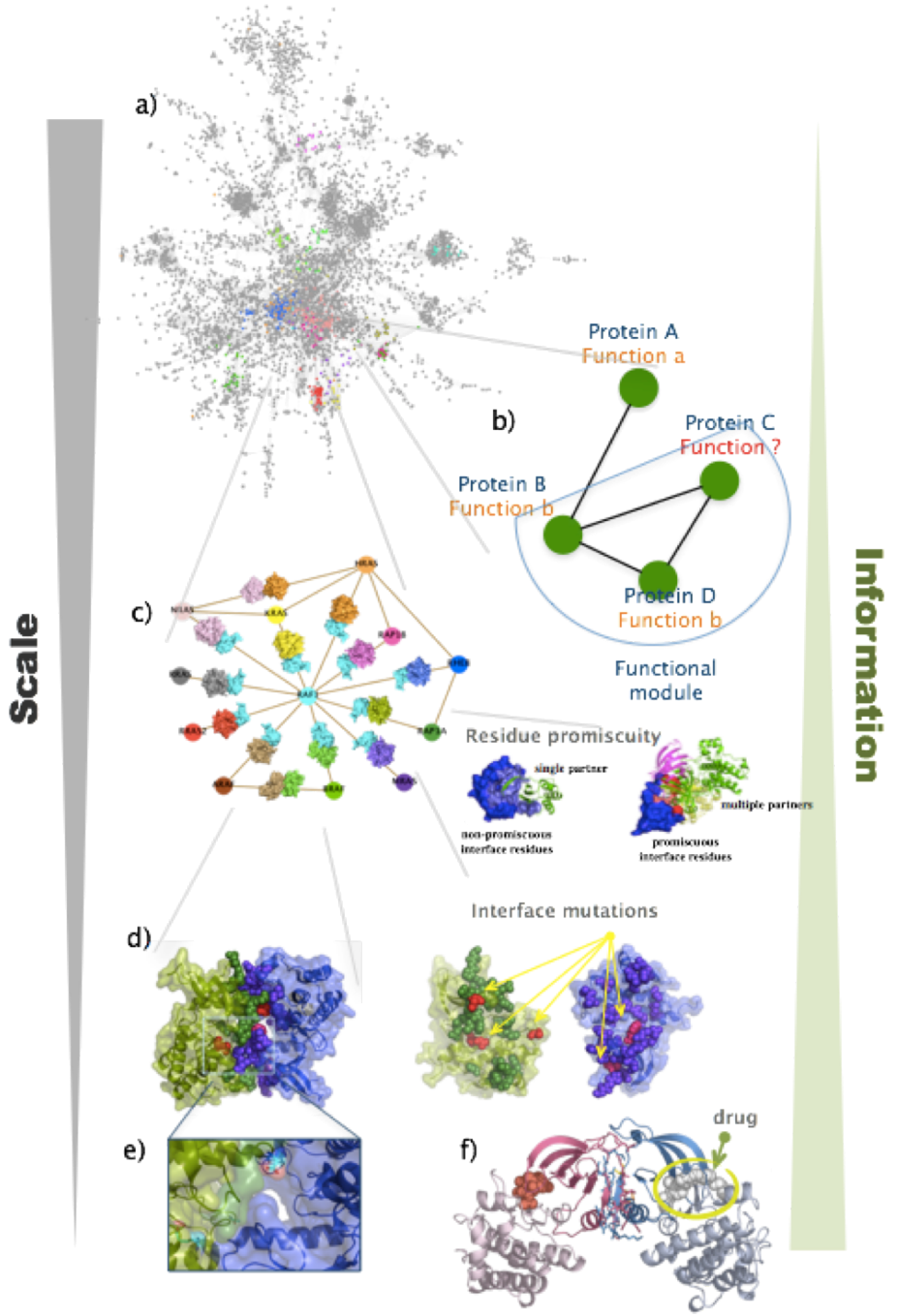

Figure 1.1: The scale of the biological system versus the amount of information a protein–protein interaction network contains.(A) The human protein–protein interaction network. (B) Guilt–by–Association principle in predicting protein functions. (C) 3D RAF1 sub–network containing information from system level to atomic level. The nodes represent proteins, whereas the edges are annotated with protein complexes. The structural information in a 3D PPIN can be used to study the promiscuity of interface residues. (D) The structure of protein complex. The interface regions are shown with shape sphere and mapped with disease-causing nsSNVs (coloured in red). (E) The atomic-level view of interface with nsSNVs. (F) A drug inhibits the binding pocket. Figure adapted and modified from Lu *et al.* [1].

Another approach is based on integrating genomic information and PPIN with the knowledge of functional pathways [16], which can be retrieved from databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [17]. These mapping has been particularly exploited for cancer studies. Genomic perturbations attributed to diseases can be mapped on biochemical pathways so as to obtain a pathway-level understanding at distinct disease states. It has been demonstrated that when cancer patients harbour genomic alterations or aberrant expression of different genes, these participate in a common pathway or have a similar effect in altering the pathway [18, 19].

These studies supported by PPINs have brought novel insight into cellular functional modules and the association between genes and diseases. By integrating PPINs with the atomic-level information (Figure 1c), one can understand more precise details on the mechanisms regulating how proteins specify their functions and how disease-causing mutations disrupt the biologically functional systems.

Pioneering studies in structure-based PPINs were done by Aloy and Russell [20] who looked into atomic details of protein interaction pairs and proposed that homologous protein pairs may interact in the same way, using the same binding interfaces. In the effort of bridging the gap between large-scale PPI determination and structural data, hybrid approaches to structure determination of macromolecular complexes have been proposed [21, 22, 23]. Integration of structural data at different resolution and reliability has been successfully used to reconstruct hybrid assembly structures that can be informative for further PPI validation studies or in designing new tailored structural investigations. These 3D network studies were initially done on the model organism yeast. Recently, in reason of the increasing number of available structures and interaction data, 3D PPINs of other organisms have also become available [24, 25], including human, mouse, drosophila, and

bacteria.

Recent studies [26, 27, 28, 29, 30] integrated protein structural information with PPINs to give the ability to implement large-scale studies on the association between cellular mechanism and protein complexes. One of the major interests in exploiting these studies is to investigate how disease-related mutations may disrupt protein functions and ultimately affect the function of biological systems. Mutations can be classified as loss of function, gain of function or neutral according to their effect on protein function. These effects can be mediated by alterations of the protein stability induced by the mutation [31, 32]. For example, the B-RAF kinase is widely mutated in cancer and the V600E mutant, recently observed also in patients with granulomatous paediatric disease [33], destabilises the inactive conformation of the kinase. This gain-of-function mutant keeps B-RAF in the active state and consequently increases the activation of another kinase ERK [33]. Mutations can affect protein function also by modifying the affinity of the protein for its partners. For example, mutations derived from Glioblastoma patients have been recently shown to destabilise the complexes of the proteins involved in the disease pathogenesis, mainly through a decrease of the electrostatic contributions to the binding energy [28]. Drastic amino acid changes at the protein interface, where a protein is in physical contact with another protein, can significantly change the binding energy of the interaction. In particular, the occurrence of mutations at the interface hot spot residues, which contribute the most to the binding energy [34, 35], are most likely to have impact on the interaction, so that the proteins would either lose interactions with the partner proteins or gain interactions with new binding proteins. Additionally, protein function can be altered by mutations occurring at allosteric sites. Indeed, allosteric mutations can disrupt or promote the binding of allosteric modulators, affect the communication pathways between the allosteric and

orthosteric sites and modify the relative proportion of inactive/active conformations [36, 37]. For example, different cancer-related mutations in kinases have been shown to involve a shifting in the relative population of the inactive/active states [38, 39].

Besides providing insight into the impact on the complex functionality of disease-causing mutations, 3D PPINs can be used for large-scale screenings of drug targets to compensate experimental drug compound screening methods [40]. A number of approaches have been developed in the recent years to exploit PPINs in drug discovery, as the cellular network and the surrounding environment are essential part of the process of efficient drug targeting and delivering. By implementing large-scale screening over protein molecular properties, one could identify new target proteins and potential binding sites of drug compounds. In particular, PPINs can be used to identify PPI inhibitors. Targeting PPIs is still one of the most challenging tasks in drug design, owing to the significant differences between interfaces in PPIs and small molecule binding sites. However, specific properties embedded in PPI interfaces and needed for partners recognition can be exploited to identify drug compound targets [41, 42, 43, 44, 45]. Indeed, there are increasing examples of targeting PPIs (Figure 1f) with drugs that can bind to transient and dynamic pockets in orthosteric or allosteric sites (for a review on the subject, see Engin *et al.* [46] and references therein). In targeting PPIs, people are also exploiting Protein-Interface Motifs to identify potential off-target drugs [47]. Databases such as Interactome3D [24] and INstruct [25] provide PPINs annotated with protein complexes and are essential to large-scale screening approaches targeting protein interfaces in drug design [48].

Another interesting use of targeting PPIN, pathways and drug action mechanisms is the identification of proteins that induce drug side effects [49]. Drug side effects are often the most undesirable outcomes from medical treatment, frequently caused by the binding

between drug compounds and off-target proteins. Adverse drug reactions (ADRs) was reported to be one of major causes of mortality and morbidity over the last decades [50]. Poly-pharmacology approaches have become popular in complementing the classical one-drug one-target paradigm [51].

Still, the bottleneck of systematic screening of binding pockets for drug compounds *in-silico* lays on the limited availability of experimental structures. Homology modelling can be used in generating 3D protein models to compensate this limitation [52, 53]. However, high quality structures are essential for binding pocket detection. Model refinement procedures can help in obtaining a more realistic structure for these drug-target binding studies [54]. However, additional challenges in protein complex prediction are in that often proteins are subject to conformational changes to attain specific binding modes. A special case is when these functional states are induced by allosteric sites signalling, generally not easy to observe experimentally [55]. Thus, conformational-change perturbations are usually not taken into account in protein complexes modelling procedures.

In the following sections, we discuss the availability and quality of PPI datasets, as well as the current state of high-throughput experimental methods for PPIs detection since they are fundamental to build a 3D PPIN. We particularly focus on recent applications of 3D PPIN, highlighting strengths and discussing limitations related to the availability of structural data for human proteins. Finally, we briefly comment on how 3D PPIN could contribute to the design of novel targeted therapies, particularly useful to the advancement of personalised medicine.

1.2 Protein-Protein Interaction Data

High-throughput experimental methods have given the possibility to build PPINs of entire organisms, which in some cases (e.g. yeast [56]) are deemed close to complete. They include detection of direct interactions by yeast two-hybrid (Y2H) assays, and detection of protein complexes by affinity purification-mass spectrometry (AP-MS). Literature curation and annotation is another useful source for PPI datasets, often extracting information obtained from small-scale experiments, such as Fluorescence Resonance Energy Transfer (FRET) or other biophysical investigations. However, these collected datasets are usually biased towards the proteins that have been most extensively studied and are not large-scale, due to constraints in the detection methods. In 2008, it was estimated that about 650,000 PPIs should occur in humans [57] and so far about one-tenth of the estimated human interactions have been observed experimentally [58]. Publicly available databases, such as HPRD [59], BioGRID [60], DIP [61], IntAct [62], MINT [63], and STRING [64], provide platforms to access PPIs curated datasets.

While the high-throughput experiment techniques are progressing to obtain complete pictures of biological systems, low reproducibility of the data has raised concern about the data quality. Braun P. [65] pointed out that the overlap of yeast PPI datasets derived from AP-MS experiments between two labs could be as low as 20%. This may be ascribed to different reasons, including the absence of the same standardised experiment protocols and biased sampling. Varjosalo and colleagues [66] demonstrated that high-throughput PPI experiments are highly reproducible when performed by two different labs if the protocols with the same standardised workflows are used. Moreover, to diminish bias sampling, they used 32 human kinases as bait proteins with a different domain composition, expressed in different tissues and involved in different biological processes. Analogously, Havugimana

and colleagues [67] generated a pipeline with stringent experiment procedures and applying computational methods to detect high-abundance components and identify functionally unrelated protein pairs. Proteomic profiles were used to assess the abundance, reducing the number of false positive interactions from protein pairs that in vivo are not expressed at the same time and cellular space. Scientists in the AP-MS field are developing experimental approaches to mitigate some of these inefficiencies, using, for example, replicated and control experiments and relative quantification to enhance sensitivity and/or by developing confidence scores to select specific protein-protein interactions [68, 69, 70]. Apart from experiment protocols, many studies also suggested the need of data standardisation [71, 72] and validation [73, 68]. The International Molecular Exchange (IMEx) consortium provides the controlled vocabularies and standardised data formats which have been adopted by major databases [74, 75]. Statistical methods [76] and structural information are also suggested for the validation of PPIs before they are deposited to the databases.

Additionally, a number of computational methods have been developed to compensate the experimental methods and expand the space of PPINs based on strategies such as co-evolution [77] and homology modelling [78] (for a recent review on computational prediction methods see Mosca *et al.* [72]). However, high-confidence PPI data with experimental evidence are fundamental to build 3D PPINs which could carry out more robust studies in disease-related mutations and drug target identification.

1.3 Building 3D PPIN

As previously mentioned, PPIN is a useful tool in identifying disease or functional relationships between proteins. Yet, system-level representations of biological processes provide

very limited information on answering crucial questions, such as how a protein recognises its partner proteins, or which region of its surface binds to its partner proteins. It requires atomic level information to understand binding mechanisms. However, mapping structural information onto networks remains a challenge due to the gap between the number of known proteins and the number of solved protein structures and some types of proteins are under-represented in structure databases, such as membrane proteins. So far, the Protein Data Bank (PDB) [79] stores roughly around 5,000 human protein structures with many of them containing only partial structures.

To tackle this problem, earlier work in developing 3D molecular interaction networks, including iPfam [80] and 3did [81], analysed the structures of protein complexes at domain level. The domain-based interactions are supported by both inter- and intra-species co-crystal structures, and they include interactions between domains belonging both to the same and to different proteins. 3did covers more than 4,000 distinct domains which is about one-third of the total number of Pfam domains [82]. Importantly, domains are the basic evolutionary and functional units of proteins. Proteins with domains in a common superfamily are considered more likely to be evolutionarily related [83]. By looking at molecular details of protein domain interactions, one could identify the domains which are functionally important in mediating PPIs.

To increase the coverage of structures in PPINs, the two recent databases INstruct [25] and Interactome3D [24] implement two different approaches both based on the use of homologous structures. One should be aware that PPI predictions using homologous structures can be of different nature. One is to use homologous protein structures of a protein pair to predict the possibility of interaction which is not detected from experimental methods. The other is to predict structure complex of a protein pair which is known to interact from

experimental methods and therefore trying to enrich with structural information in the available large-scale screens. The following discussions are based on the second strategy to predict the protein complexes. The pipeline of INstruct to generate a 3D PPIN starts from binary interaction datasets from different publicly available databases. Each interacting pair is then annotated with the corresponding co-crystal structure if available, or with co-crystal structures of homologous proteins. It should be noted that the resulting structural annotation of protein pairs with homologous co-crystals is only approximate. The database provides 3D PPIN data of human and six other most studied model organisms, where human 3D PPIN contains 6,585 interactions between 3,627 proteins. A different strategy was used for Interactome3D, where the structural coverage of human PPIN was increased by modelling interacting pairs with missing structural data using Modeller [84]. This provides a more precise representation of interface regions of interacting protein pairs. Interface residues are identified by calculating the distance of residues from protein pairs. The database provides human 3D PPIN which contains 6,473 interactions between 4,239 proteins.

One may also use predicted protein structures obtained from reliable resources to compensate the limitation of solved protein structures. A recent project Genome3D [85] integrates UK-based structural resources, including Gene3D [86], FUGUE [87], and four other structural prediction resources [15, 88, 89, 90]. The aims of this project are to provide biologists a platform to compare the predictions from those resources which were developed with different algorithms, and to choose the prediction outcomes which are more reliable. Those predicted structures could help to expand the size of 3D PPINs.

To identify or even predict the proteins that can be acting together, one could use available structural information. A study by Kar *et al.* [91] , for instance, looked at the structural

features of cancer proteins. Cancer proteins are well known to involve biological processes related to, for example, DNA repair and cell growth. In this study, ten functional pathways were selected according to the Cancer Cell Map (http://cancer.cellmap.org/cellmap). Each protein pair in a given functional pathway is annotated with structural information obtained by running PRISM [92], a software to explore the known protein-protein interface binding modes and predict analogous cases. PRISM can predict the protein interaction by searching interface with similar backbone geometry in the interface library. In this way, co-crystal structures are not strictly required to build the PPIN of the pathway. This method is beneficial for smaller scale studies with interests in proteins in particular functional pathways and save computational time to construct a 3D PPIN of an entire proteome.

To summarise, in this section we have reviewed some of the recent approaches to generate 3D PPINs. As it will be shown in the following sections, building increasingly complete 3D PPINs is essential to interpret biological systems, provide insight into complex cellular mechanisms and rationalise genotype-to-phenotype relationships.

1.4 Protein Interaction Interfaces

The ability of proteins to recognise and bind their partners is essential to biological processes. Protein interfaces, where proteins have physical contact with their partner proteins, are believed to embed crucial properties which mediate PPIs. Both experiments and computational analyses have shown that protein interfaces mediate PPIs through specific molecular properties, including sequence motifs [47, 93], backbone geometry [94], residue types [95], interface hot spots [34, 96], and correlated changes in the two interfaces.

Precise atomic coordinates of protein complexes in 3D space are fundamental for studying the physical and chemical properties of protein interfaces. The protein structure repository Protein Data Bank [79] documents more than 90,000 protein structures in complexes (data from April 2014) determined from different experimental observations, including X-ray crystallography, Nuclear Magnetic Resonance (NMR), and electron microscopy. However, these crystal packings of molecular units may not always represent natural protein interactions. Biological relevant interactions may be sacrificed in crystal state in order to minimise global free energy. This could result in unspecific macromolecular interactions.

Hence, the atomic coordinate files in PDB database are classified as either biological assembly or an asymmetric unit. The former represents the biologically-relevant complexes which are either provided by the authors, validated using the software PISA [97], or both. PISA calculates physicochemical properties of a micromolecular structure to estimate the stability of the macromolecular complex. These properties include free energy of formation, gain in solvation energy, hydrogen bonds and saltbridges across the interface, and hydrophobicity. The interfaces obtained from biological unit coordinate files are more reliable to be the biological interface. The asymmetric assemblies contain either one biological assembly, a part of a biological assembly, or a combination of biological assemblies depending on the crystallisation conditions for forming a crystal. Therefore, the interfaces identified from the coordinate files of the asymmetric unit may not be protein functional interfaces.

Each interface of a protein is comprised of several discontinuous patches. Two estimates have been commonly used to define the interface regions. The first approach is to calculate the distances between residue pairs from two proteins in a co-crystal structure [24, 44]. Two residues are considered as interacting if their distance is within a pre-defined

threshold. This value may depend on the group of residue atoms that is used to calculate the inter-residue distance. Typical threshold values are 4-5 Å on the distance between any pair of atoms from the two residues [98, 81] or 9 Å on the distance between C^{α} or C^{β} atoms [44]. The sum of atomic van der Waals radii + 0.5 Å is another frequently used distance threshold [99]. More sophisticated criteria use different thresholds for different types of interactions (e.g. hydrogen bonds, salt bridges and van der Waals interactions) [24]. The second approach compares the solvent accessible surface area (SASA) of the protein complex with that of the single components to evaluate the area buried upon complex formation (interface area). For example, residues can be considered as part of the interface if their total or side chain buried SASA is larger than 0-1 $Å^2$ [100, 101]. The calculation of the buried area can be implemented in automated approaches such as POPSCOMP [102]. Pre-compiled values for PDB complexes are also available from databases such as 3did and PIBASE [103]. For protein pairs where the structure of the single proteins is available but not that of the complex, molecular docking methods, such as FiberDock [104], can be used to predict the interface regions.

In one of the leading studies on biologic structural networks, Kim and colleagues [105], classified hub proteins into two groups according to the number of their interfaces: multi-interface hubs and singlish-interface hubs. The two groups were shown to have different evolutionary properties. In particular, only multi-interface hubs turned out to be significantly more essential and slow-evolving compared to the average. Since early studies considered instead all hub proteins to be essential [106, 107], this shows that the integration of sequential and structural information on protein interfaces can increase the precision in identifying functionally important proteins within biological systems.

The importance of interfaces for protein functions and biological processes was further

confirmed in recent studies by looking at the occurrences of disease-associated mutations [27, 108], where interface regions were found to be enriched in disease-causing mutations. This implies that a residue change at these region is more likely to disrupt protein functions and lead to diseases. In a follow-up study by the same group [109], further annotations were given to the mutations mapped on 3D PPIN. Dominant truncating disease mutations were found to have different pattern to other classes of mutations with no preference occurring at interface regions. Moreover, recessive mutations co-localised on the same interface, showed the tendency to cause the same disease.

Increasing the level of detail in the description of protein interfaces further highlights the importance of structural properties in determining the protein function. For example, promiscuous binding sites, which are essential for hub proteins to interact with many different partners, can be identified by mapping interactions with multiple partners on the protein surface. Promiscuous sites have been shown to possess specific properties in terms of amino acid composition [110, 111], solvent accessibility [112], packing [99] and conformational flexibility [113, 114], mainly related with their increased capacity to adapt to different partners. The biological relevance of promiscuous residues is further confirmed by a recent study from our laboratory [113], showing that they are less enriched in nsSNVs. This finding suggests that residues in promiscuous positions have a reduced tolerance to genetic variations, related to the necessity to preserve their binding poly-valence.

Although key challenges remain in performing interactome-scale studies on protein interface regions due to the low availability of protein structures, computational approaches may help in overcoming this limitation. Gao and Skolnick [94] found that many protein complexes have similar interfaces even if the overall structure of the single components is different. Thus, they argued that even though there are only a small fraction of pro-

teins with solved structures, the protein interface library is close-to-complete. This is the fundamental idea behind protein binding pocket searches [115, 116] and PPI predictions that use interface structure similarity scoring [78, 92] in order to increase the structural coverage of 3D PPIN.

1.5 Molecular Properties of Interfaces

Based on complex composition, affinity and lifetime, protein complexes can be classified into different types: homo-oligomeric or hetero-oligomeric, non-obligate or obligate, and transient or permanent complexes [117]. Protein interfaces play a role in molecular recognition and in determining the type of molecular interaction. Numerous studies have explored the functional, evolutionary, and physicochemical properties of interfaces. Identifying these properties is crucial for understanding not only how proteins determine biological functions but also how disease-related mutations can disrupt cellular systems.

Early studies in characterising interface molecular features suggested that interfaces exhibit distinctive features when compared to other parts of proteins. They are generally flat, enriched with hydrophobic residues, most often forming part of helical secondary structures [101], and containing a number of charged groups [118]. The residues at interfaces are also more accessible than other solvent exposed residues and have fewer intra-molecular contacts. More recent studies [119, 120] further examined and characterised interfaces of different types of protein complexes, including homo-dimeric, hetero-dimeric and transient protein complexes. Interfaces of heterocomplexes were found to be more planar and less hydrophobic than homodimers. Whereas, transient complexes with smaller contact areas are generally more conserved, planar and polar. Those with larger contact areas (>1000

Å^2) often undergo conformational changes upon forming/breaking interactions [120].

1.6 How 3D PPINs Contribute to Biology and Biomedical Sciences

Advances in genome sequencing techniques and large-scale genome sequencing projects, including the 1000 Genomes Project [121] and the International HapMap Project [122], are boosting the amount of available gene variation data. Databases, including dbSNP [123], OMIM [124], COSMIC [125], HGMD [126], and Exome Variant Server (http://evs.gs.washington.edu/EVS), provide online interfaces to easily access gene variation datasets and serve with different purposes. The challenge is now to develop methods and tools to extract useful information from this increasing amount of data. In particular, it will be essential to understand what information those mutations are carrying, how we can use those gene variation data to unravel the underlying cellular mechanisms, and how disease-causing mutations lead to diseases [127, 28]. To answer these questions, atomic-level information of protein complexes is crucial to provide the biological features of disease-related proteins and disease-causing mutations. Moreover, by implementing large-scale studies with 3D PPINs, which contain information from system level to atomic level, one may find explanations for the effects of these mutations. For example, Kar *et al.* [91] implemented human structural protein interface network (iSPIN), in which the edges represent binding interfaces between protein pairs obtained from either known or predicted structures from PRISM. Their results show that cancer proteins tend to be hubs in the network. The mutations occurring at cancer protein interfaces, disrupting protein bindings and causing loss of protein functions, have greater impact on biological systems. The binding interfaces of

cancer proteins were also shown to have specific properties; they are significantly smaller, more planar, less compact and less hydrophobic than interfaces in non-cancer proteins. These specific features of cancer protein interfaces may be used for the identification of new targets and drug candidates in cancer therapies. The results demonstrated how 3D PPINs can enable more comprehensive studies in biological systems with informative outcomes. Besides, 3D PPINs can be effectively used in high-throughput screening for drug targets. Indeed, PPIs are promising druggable targets since their selective inhibition [128] can be used to regulate particular functions in biological systems. 3D PPINs, by representing in a synthetic and comprehensive way the associations between a target protein and its partner proteins, are an invaluable tool to understand the possible effects of inhibiting the binding interfaces of the target protein. All these examples show how 3D PPINs can give an essential contribution to Biomedical Sciences.

1.7 Conclusions

The studies of PPINs have progressed from system level representations of biological systems to more detailed representations annotated with atomic information. The integration of data from the currently available biological databases has proven to be essential to carry out comprehensive studies on cellular mechanisms and the causes of their disruption. 3D PPINs analysis has been used to study the role of disease-causing mutations. So far, despite the general agreement on the propensity of disease-related mutations for protein interface regions, the general characteristics of these mutations and how they affect biological functions remain challenging. Still, 3D PPINs analysis is very important to unravel the features of these mutations and their impact on protein functions. In particular, the identification of properties specific to pathogenic mutants is crucial to develop methods

for the prediction of disease-related mutations. Besides, 3D PPINs analysis can also help biologists to effectively search for possible targets for disease treatment as PPIs are ideal drug targets [128] to regulate biological functions and the structural information in 3D networks provides the guidance for the design of drug compounds in the early stage development. 3D PPINs provide fundamental materials for the screening of off-target PPIs. The use of these preliminary investigation strategies could effectively reduce time and cost in drug development. The increased understanding of the pathogenic mechanisms triggered by disease mutations, and of the activity of drug compounds in the cell, combined with personal genome sequencing profiles, will promote the development and delivery of more effective personalised clinical treatments in the foreseeable future.

The studies using 3D PPINs is a relatively new research field. The bottleneck in the generation of complete 3D PPINs lays in the limitation of available experimental data on genome sequences and protein structures. With the rapid progressing of experimental technologies and bioinformatics approaches, more biological data will become available to provide a more complete view of biological systems. In the coming years, the importance of protein binding mechanisms and the general characteristics of disease mutations will be better understood. Therefore, it will be possible to develop more sensitive predictors of disease-causing mutations, resulting in further progress towards effective personalised medical treatments.

Chapter 2

Human Gene Variants

2.1 Introduction

Different types of genetic variations and their molecular effects on cellular functions will be presented in this chapter. Recent advances in genome sequencing technology have largely enriched the availability of human genome sequence data and enabled, numerous studies on genotype-phenotype associations. The main source of the genome sequence data is large consortia, and in particular, the 1000 Genomes Project [121], the international HapMap Project [122], the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/), and Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium [129]. Additionally, consortia such as ENCODE Project [130] aim ambitiously to map the human genome landscape and to outline the underlying functional elements, which span from coding and non-coding transcripts to the marking of accessible chromatin and protein binding sites. Presently, the main challenges in the field of genomic research are twofold: Firstly, seeking the rules and patterns from this enormous and complex data archive, and secondly, understanding

the underlying mechanisms linking genetic variability to specific phenotypic traits.

The majority of genetic variations are harmless and contribute to the differences between individuals, while some of them can affect and alter cellular function using a variety of mechanisms. Human genetic variants, including structural variations and Single Nucleotide Variations (SNVs), have been intensively studied. The former refers to all genomic changes that are not single nucleotide substitutions, including insertions/deletions (indels), duplications, inversions, and Copy Number Variations (CNVs) [131]. The study of structural variations mainly focuses on the copy number variations which are often involved in altering genomic structure and gene expression levels. Whereas, SNVs, being the most ubiquitous form of genetic variations in human genome, have attracted a strong scientific interest lately, with focus on the missense mutations that substitute amino acids on protein sequences, as they can have a direct impact on protein structures and functions.

This chapter will present the current state of both experimental and computational methods related to the study of genetic variations. I will first introduce the different classes of genetic variations, followed by a description of the methods implemented to identify the variants in the human genome. The effects of genetic variations will also be discussed in terms of their molecular impact on the cellular function.

2.2 Classes of Human Genetic Variation

Human genetic variants can be defined as common or rare by the Minor Allele Frequency (MAF) in an observed population. Variants with MAF more than 5% are often referred as common variants, whereas rare variants are defined as having a frequency of less than 1% [132]. Genetic variants can also be broadly divided into SNVs and structural vari-

ants. Structural variants includes insertions/deletions (indels), duplications, inversions, and CNVs. The distinction between indels and CNVs is in the length of the variations. The indels are generally defined as having a length of insertion, deletion, or duplication less than 1kb, whereas CNVs are larger than 1kb [133]. Different types of genetic variants have different roles in biological systems and can affect human health through disparate mechanisms. Structural variants, for instance, have been suggested to be ubiquitous and frequently related to the rearrangement of genes which alters the regulation of nearby genes [134], while SNVs are the most common genetic variations between individuals and are mostly neutral [135].

The classification of genetic variations is more complicated and difficult when looking at cancer genomes. Cancer genetic variations can be classified as either driver, or passenger mutations, according to their functional effect on proteins. Driver mutations provide a selective growth advantage to the cancer cell, whilst, passenger mutations are present in cancerous cells but do not promote growth [136]. Currently, to distinguish between these two types of cancer mutations relies on experimental methods, which will be discussed in the following section.

2.3 Detection of Disease-related Genetic Variations

It is commonly agreed that evolutionary forces influence genetic variation. The theory of natural selection has raised many questions yet to be answered in modern biological and biomedical science. Questions raised include the mechanisms of neutrally evolving polymorphisms, the role of genetic variations in determining phenotypic traits, and the properties and mechanisms of mutations which alter the fitness of humans.

One classic example of human genetic studies is the immune resistance to malaria of the sickle cell mutation carriers. Malaria was one of the major causes of mortality in human history, and still is in certain part of the world [137, 138]. It has been suggested to be the evolutionary driving force behind sickle cell anaemia disease, a recessive disease [139, 138]. People who carry the sickle cell allele at the human haemoglobin β locus have been reported to have a higher chance of surviving malaria [140, 141]. In particular, the sickle cell heterozygotes, who carry only one copy of sickle cell mutation, have even greater resistance to malaria than the homozygotes [140]. The homozygotes, inherited from both parents, also suffer sickle cell anaemia. This is an example of balanced polymorphism where the individuals carrying both versions of a gene have better ability to survive or adapt to the environment [142].

Studies that involve the detection of genetic variations for specific phenotypic traits require genome sequence data for comparative analysis of genes between different populations, most typically between patients and healthy individuals. Two commonly used experimental approaches to detect genetic variation include sequence-based and array-based. Human genome sequence data of healthy individuals from the genome sequencing project, such as the International HapMap Project, have often served as reference sequences to be compared with the sequences of patients to identify and characterise the variants which are responsible for the specific disease being analysed. Moreover, genomic microarray technology, has also been used to determine the genes or genetic variations that are associated with specific diseases. Many genotyping arrays have been designed based on the HapMap genome sequence data. The probes of the microarray allow for the effective detection of indels and CNVs in the human genome and therefore are often used as a clinical diagnostic tool for the diseases that are caused by CNVs, such as autism disorders [143]. The

downside of genomic microarray is that it heavily relies on frequency information from the healthy control to reduce the background noise [134].

Genome-Wide Association Studies (GWAS) are the most widely used statistical approaches to detect disease-related variants in large-scale. It is particularly effective in identifying genes or variants that are associated with complex diseases such as diabetes, breast cancer and prostate cancer. A growing numbers of novel loci have been identified and assessed for statistical associations with specific diseases by GWAS over the last few years [144]. However, debates about GWAS methodology mean that the implementation and interpretation of results should be done with caution particularly the selection of case and control datasets (see [145] for more details on this topic), and the under-representation of rare variants due to limited sample size [132]. The underlying mechanisms of the association between genes and diseases remains largely unknown.

Many bioinformatics tools have been developed to predict the effects of exonic single point variants which substitute amino acids on protein sequences: SIFT [146], PROVEAN [147], PolyPhen2 [148], Panther [149], SNAP [150], i-Mutant [151] and many others. These prediction tools utilise different approaches to measure the impact of variants on protein functions. The most often used parameter in discriminating disease-related variants from neutral variants is conservation scores because functionally important regions are generally conserved. Some methods, such as PolyPhen2 and SNAP, increase the precision of the prediction by including structural information, such as secondary structure prediction and the physico-chemical properties of the involved residues. The method SAAP [152] further mapped mutations onto structures to measure structural effects and predict resulting phenotype of mutation using machine learning methods with a list of predefined features. Other approaches, such as calculating $\Delta\Delta G$ values to measure the impact of the amino

acid substitutions on the stability of the overall protein structures, are also used. Those tools, mentioned above, often use the variation datasets obtained from UniProt, OMIM, or dbSNP to train their methods. The prediction tools have reached fairly high accuracy with germ-line disease variants [153].

Identifying cancer mutations, however, remains a challenge, particularly, the identification of driver cancer mutations. These mutations have been suggested to play an important role at the early development of cancer [136]. The identification requires many cancer cell sequencing samples in order to detect highly mutated genes and recurrent variants. To reduce the number of targeted genes and variants for further investigation, computational methods have been developed specifically for cancer driver mutations, including MutationAssessor [154], MuSiC [155], CHASM [156], mCluster [157], CanPredict [158].

2.4 Molecular Mechanisms of Disease-related Mutations

Different types of disease-related mutations affect human fitness through a number of possible mechanisms, most of these remain to be elucidated. Among those mutations, CNVs and missense variants are currently the most studied mutation types, as they were found to be the most common forms of disease-related variations.

CNVs have been shown indeed associated with many pathogenic phenotypes. Particularly, they are often suggested to be involved in developmental disorders, including intellectual disability (ID) and autism [159]. So far, CNVs have been reported to affect the amount of gene expression by altering the numbers of genes in the genome or by occurring in the gene regulatory elements to up-regulate or suppress gene expressions [160, 161, 162]. In fact, the genome of healthy individuals would also contain several large size CNVs [163].

However, the underlying mechanisms from the evolutionary point of view, in terms of their actual functional pressure in the transformed genome remain to be discovered.

In contrast, SNVs are relatively well documented in many databases. Among different types of SNVs, missense variants have been most studied due to the unique characteristic that they can directly affect gene products by substituting amino acids in protein sequences. In the following, and for the rest of the presented work, missense variants will be referred as non-synonymous Single-Nucleotide Polymorphisms (nsSNPs).

The growing number of available crystal structures of protein complexes offers the unique opportunity to study the structural features of nsSNPs. Many studies have shown that drastic change of amino acids may destabilise or alter protein structures and function. One example is shown in the study of patients with aldosterone synthase deficiency (ASD) type I [164]. The mutation R374W in CYP11B2 protein causes the loss of hydrogen bonding activity around residue R374. Hydrogen bonds is one of the most important elements for protein stability. The mutation R374W consequently abolish enzyme activity of the protein. nsSNPs can also affect biological systems by altering cellular interactions of the mutated proteins. The aggregation of amyloid fibrils in neurodegenerative disease is a classic example. The mutation D23N in amyloid β-protein was found to disrupt the local structure of the protein and alter the hydrophobic core region to expose to the water phase [165]. This consequently triggers the aggregation of amyloid β-proteins. The results from recent studies [26, 27, 28, 29, 30] have also implied the importance of protein complexes interface regions. Variants occurring at protein complex interface region are likely to be associated with diseases. Moreover, genetic variations can affect protein function through other mechanisms not directly occurring at ordered sites of protein complexes. Post-translational modifications (PTMs), for instance, are often required for changing the state

of proteins to perform specific functions. Mutations occurring at or close to PTMs may result in a dysfunctional protein or keep the protein constitutively in a specific state. Additionally, protein function can also be altered by mutations occurring at allosteric sites. This consequently affect the communication pathways between the allosteric and orthosteric sites, and modify the relative proportion of inactive/active conformations [36, 37].

2.5 Intrinsic Disorder

The focus of IDRs in this project is on the association between human mutations and disordered structure regions and the features of the mutations occurring at these regions.

To fold globular-like compact structures had long been considered as the profound paradigm for proteins to carry out their biological roles. It was only a decade ago, intrinsically disordered proteins and Intrinsically Disordered Regions (IDRs) began to come to the attention. Numerous proteins, particularly in Eukaryotes [166, 167], were found to be lacking a unique fold, either partially or entirely, and play nevertheless essential roles in regulatory and signalling processes [167, 168]. In fact, proteins can exist in four different structure states: a) compactly folded, b) compactly folded domain(s) with disordered regions (2.1), c) compact but disordered, and d) extended intrinsically disorder. These can be grouped into three broad states: ordered, collapsed (molten globule) and extended [169, 170, 171]. The different structure states can undergo transitions from disordered to ordered upon binding to their interaction partners [172]. The strong interest in IDRs lays with their sequential and functional features, which includes the related sequence signatures, the biological roles, binding mechanisms, their specific interaction partners, the association with diseases, and

Figure 2.1: Partially intrinsically disordered structure - CD23 (PDB 1t8c) [173].

their involvement in druggable sites.

Thus far, IDRs have been studied mostly based upon the amino acid sequence information, due to the fundamental constraints of experimental methods. In particular, crystal-structure analysis requires stably folded-state structures for electron density quantifications. This has limited the available IDR structures [174]. To compensate for this limitation, computational methods have been implemented to predict the disordered segments on protein sequences. Boosted by the recognition of the IDR important biological roles, more than 50 prediction tools are currently available, including MoRFpred [175], DISOPRED2 [176], PONDR-FIT [177], FoldIndex [178] amongst others (see [179] and [180] for the comprehensive introductions of IDR prediction tools). The attributes adopted by

those tools are based on the currently known biochemical properties of IDRs, often including amino acid composition, hydrophobicity, secondary structure, solvent accessibility and other physicochemical properties.

Although there has been significant progress in this area during the last decade, the following properties of IDRs remain challenging to be fully understood. Firstly, the molecular mechanisms of functioning for IDRs remain unclear. Current studies are mostly relying on amino acid sequence information. The database ELM [181] provides known experimentally validated motifs and a tool SLiM to identify putative linear motifs. The linear motifs are composed of 3 to 11 contiguous amino acids and enriched in unstructured or disordered regions of proteins. They often play an important role in regulating processes. These documented short linear motifs could help to understand the molecular and sequence features of functional segments at disordered regions. Just as ordered protein regions have preferential physicochemical compositions of amino acids to form stable 3D structures, disordered regions have also been reported to have compositional bias towards polar and charged residues and to be lacking bulky hydrophobic ones [182, 183]. Still, this gives very limited understanding of the association between the disordered structures and their functional mechanisms.

Tools, such as MoRFpred, have been developed, using multiple sequence alignments, to predict the functional segments on disordered regions called molecular recognition features (MoRFs). However, these tools were trained and assessed with a very small number of experimentally characterised disordered structure information. According to the database DisProt release 6.02 (May 2013) [184], which documents currently known intrinsically disordered proteins, only a total number of 694 disordered proteins and 1,539 disordered regions have been recorded in the database.

Moreover, different types of IDRs may exhibit different biochemical properties, adopt different molecular mechanisms, and associate with different functions [185, 186]. One interesting example is the HSP33 protein which undertakes ordered-to-disordered transition as a defence mechanism to protect proteins from toxic aggregation and oxidative stress-induced cell death [187]. This ordered-to-disordered transition differs from the commonly known principle that IDRs are activated by post-translational modifications or phosphorylation and undertake disordered-to-ordered transition or remain disordered to determine their functions. The variety of IDRs increases the level of difficulty in the predictions and related studies.

The study of IDRs in the context of evolutionary conservation and implications for drug design will not be discussed here as the focus of this project is on the human genetic mutations. The occurrences of mutations at disordered regions that change the properties mentioned above may have a dramatic impact on protein functions, such as the transition between disordered and ordered states, and post-translational modifications. Thus far, no systematic study has been presented for the role of IDRs in human mutation data and to identify possible associations between IDRs and diseases. The computational methods do not have sufficient sensitivity to predict the impact of mutations at disordered regions (see chapter 3 section 3.4.7). This limitation is a result of the software being developed using parameters which are based on the structural and sequence properties of protein ordered regions. The diversity of disordered regions has also increased the level of difficulty in predicting the impact of mutations at this region.

By identifying the common properties of disease-related mutations in disordered regions, we hope to see if the common biochemical properties suggested by previous studies are indeed essential in maintaining disordered structures and their functions. The mechanisms

of determining biological functions and the impact of mutations at IDRs is presently insufficiently documented and requires further exploration. A systematic study of human disease-related mutation data at disordered regions will be presented in the next chapter.

2.6 Conclusions

Different types of genetic variations and their molecular effects on cellular functions were presented in this chapter. Genome sequencing projects have improved our understanding of genetic variations and their effects on cellular activity. However, the mechanisms of most of the variations need to be further explored to complete our understanding of their impact on diseased and healthy states. Sufficient biological data, including human genetic data, structural and functional information enables more comprehensive studies on missense mutations, particularly, for cancer mutations. Many computational methods have been developed to predict the effects of germ-line disease mutations. The properties and the underlying mechanisms of those variants needs to be better characterised. My work in the next chapter highlights some of the progress made in this area.

Chapter 3

Computational Analyses of Disease-related Variants

3.1 Introduction

Understanding how genetic variations contribute to phenotypic diversity and influence individual's fitness is of great interest in the field of biomedical research. Large-scale genome sequencing projects, such as the 1000 Genomes Project [188], have been providing an invaluable resource of human genetic variation data. Single-Nucleotide Polymorphisms (SNPs) are the most prevalent form of genetic variations with an estimated occurrence in the human genome of more than 11 million [132]. The majority of these are harmless and neutral, and mainly responsible for subtle differences between individuals, including the response to the environmental factors and the physical appearance.

In addition to the disease-related SNPs occurring at functional genomic regions, recent studies have focused on single amino acid variants. These substitutions of amino acids are

called missense variants or non-synonymous SNPs (nsSNPs) and can have a direct impact on protein structures and ultimately alter protein function or disrupt the interaction with partner proteins.

Statistical methods, such as genome-wide association studies, have been used to study the patterns and frequencies of variants between different populations or species for the purpose of identifying genes and variants that contribute to specific phenotype traits. [189, 190, 191, 192]. However, the output of these studies requires further validation using experimental methods, such as comparative analysis of hybridisation intensities on micro-arrays [133], transgenic and knock-out methodologies [193]. Moreover, statistical analyses associating variants and diseases provide limited information on the underlying biological mechanisms that link variant occurrences to diseased cellular states.

Numerous computational methods have also been developed to predict the effects of nsSNPs on protein functions, such as PolyPhen2 [148], SIFT [146] and SNAP [150] amongst others. Sequence conservation scores are commonly used as an attribute to predict the impact of nsSNPs. Indeed, functionally important residues are generally conserved throughout evolution and nsSNPs on sequence conserved regions are more likely to be deleterious. PolyPhen2 and SNAP further include structural information, including structure-derived constraints and physicochemical attributes of the variant changes, to predict the effects of SNPs. In general, these methods have a moderate accuracy in distinguishing germ-line disease SNPs from neutral SNPs. However, those prediction methods are comparatively less sensitive and accurate for somatic cancer SNPs [194]. This may due to the fact that cancer variants exhibit more diverse and variable properties than germ-line disease SNPs and that these properties are not well characterised and defined as yet. By comparing germ-line disease and somatic cancer SNPs at the atomic-level, we may find more sophis-

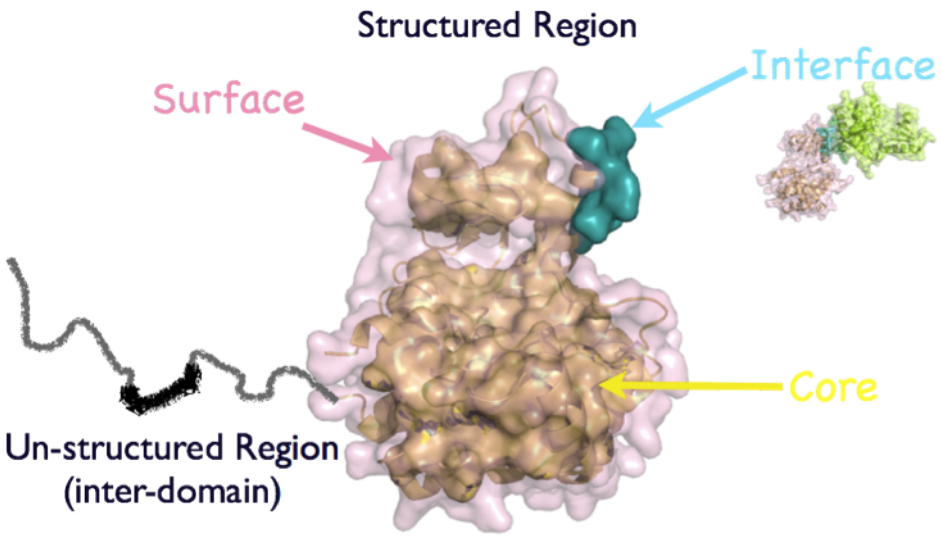

Figure 3.1: Defined protein regions. Each protein in the human 3D PPIN is dissected into four regions if possible. The domain region (structured region) includes surface, interface, and core, while the disordered region is outside of the domain region and defined by the predictor DISOPRED.

ticated features to distinguish the different types of disease-related variants. This would be essential to develop more robust prediction methods, as well as to understand how disease-related SNPs affect protein functions and lead to disease states.

3D PPINs have been successfully used in a number of recent studies to validate or predict novel Protein-Protein Interactions (PPIs) [23], characterise protein binding interfaces [113, 112, 195], and study human mutations on protein structure complexes [27, 26, 108, 28]. Here we present a large scale study of disease-related nsSNPs using 3D PPINs. We built up our own human 3D PPIN to maximise the number of 3D structures mapped onto the available human interaction data. We aim at characterising the structural features (surface, interface, core and disordered) of disease-related nsSNPs, including germ-line disease nsSNPs and somatic cancer nsSNPs as obtained from the Online Mendelian Inheritance in Man (OMIM) [124] and COSMIC [125] databases, respectively. The structural and functional information in the constructed 3D network enables a more detailed insight into

molecular features of disease-related nsSNPs. The selected nsSNPs were classified by their occurrences in the defined protein region classes (surface, interface, core and inter-domain disordered) (Figure 3.1), since nsSNPs in different classes can affect protein functions via different mechanisms. SNPs occurring at the protein core, for instance, can alter or destabilise the structure, while SNPs occurring in the proximity of a phosphorylation site at the surface of a protein can possibly modify inactive/active conformation equilibria. Recent examples of human genetic variation studies have reported the enrichment of disease-related variants at protein interface regions, where proteins have physical contact with other proteins [27, 26, 108, 28]. These genetic variations may disrupt protein functions by altering the ability to bind to their partner proteins. The same enrichment was observed in our collected disease-related nsSNP datasets (Figure 3.2).

In recent years the functional importance of protein disordered regions has come to notice, so that they were included in our nsSNPs classification. Indeed, a large number of proteins in eukaryotes have been found to be entirely or partially disordered [167, 168]. Disordered regions can have a critical role in molecular recognition and protein binding. In particular, they can undergo major structural changes upon binding to partner proteins [196] and exhibit disorder to order transitions as those observed in Molecular Recognition Features (MoRFs). Many studies on disordered regions have reported their essential role in cell signalling and their implication in cancer [197, 198]. Their lack of a defined structure raise a question about their tolerance to genetic variants and the underlying mechanisms associated with disease. Therefore, the analysis of disordered regions included here can give an insight into their molecular and functional features in comparison with structured regions.

In the following, we will first analyse the distribution of different types of nsSNPs at dif-

ferent protein regions ("**3.2.1** Enrichment Analysis of Disease-related nsSNPs"). Then, the relationship between nsSNPs occurrence and structural constraints will be investigated ("**3.2.2** Structural Features of Disease-related nsSNPs"). In particular, we will compare the propensity of nsSNPs at flexible and stiff regions, as defined by their secondary structure annotations. Moreover, the relative frequency of drastic and moderate amino acid changes induced by nsSNPs will be analysed. At last, the functional specificity of disease-related nsSNPs ("**3.2.3** Functional specificity of nsSNPs") and their co-localisation on protein interfaces ("**3.2.4** Co-localised disease-related nsSNPs") will be investigated.

3.2 Results and Discussion

An automated pipeline has been developed in this study to construct 3D PPINs (see details in Chapter 4). The human 3D PPIN generated from the pipeline was composed of 8,249 proteins with 39,387 interactions. Each protein in the network was annotated with structural and functional information. Two disease-related nsSNP datasets were obtained from OMIM and COSMIC databases that document Germ-line Disease (GD) nsSNPs and Somatic Cancer (SC) nsSNPs, respectively. In the following, GD nsSNPs will be referred to as nsSNPs^{GD}, while SC nsSNPs will be referred to as nsSNPs^{SC}. A total of 12,761 nsSNPs^{GD} and 202,719 nsSNPs^{SC} were mapped onto the 3D network annotated with 3D information. A control dataset was also obtained from dbSNP [123]. After filtering by omitting known disease-related nsSNPs and mapping onto the 3D network, the control dataset contains 461,674 nsSNPs with 3D information. This dataset is referred to as common nsSNPs (nsSNPs^{C}) in this study as it includes nsSNPs that are not known or not yet known as disease-related.

	Surface	Interface	Core	Disordered	Mapped	Unmapped
Common nsSNPs ($nsSNPs^{C}$)	42,898	15,086	43,549	63,755	165,288	315,474
Mean(number of residues) ± SEM	126 ± 1.2	46.2 ± 0.5	135 ± 2.6	179.8 ± 3.2	-	-
Germ-line disease nsSNPs ($nsSNPs^{GD}$)	2,516	1,155	2,073	954	6,698	6,063
Mean(number of residues) ± SEM	157 ± 3.4	58.4 ± 1.5	186.2 ± 9.3	175.7 ± 8.6	-	-
Somatic cancer nsSNPs ($nsSNPs^{SC}$)	21,581	8,420	21,575	25,348	76,924	125,795
Mean(number of residues) ± SEM	127 ± 1.2	47 ± 0.5	140.4 ± 4.2	181.5 ± 3.2	-	-

Table 3.1: Number of nsSNPs mapped on protein complexes of Human 3D PPIN. The number of nsSNPs from each class and the average size (number of residues) of the protein regions are listed. The standard error of the mean (SEM) is also reported. The column "Mapped" reports the total number of nsSNPs from each data group that were mapped on the proteins of the 3D PPIN. The number of nsSNPs from each data group that were not possible to be mapped on the human 3D PPIN is reported in the column "Unmapped".

3.2.1 Enrichment Analysis of Disease-related nsSNPs

As mentioned previously, nsSNPs occurring at different protein regions can affect protein structures and functions through different biological mechanisms. Therefore, each collected nsSNP data group was classified into four classes according to the region of their occurrences and annotated as $nsSNPs_{surface}$, $nsSNPs_{interface}$, $nsSNPs_{core}$ and $nsSNPs_{disordered}$ (see Methods "Interface Identification"). The results obtained on nsSNP classes were compared within each nsSNP data group with statistical tests. The number of nsSNPs in each class and the average sizes of the regions are listed in Table 3.1. The enrichment of a given nsSNP class in a protein region was measured by the abundance of the nsSNP class at that region normalised by the overall abundance of the nsSNP class in the whole protein (see Methods Formula 3.2).

As expected, the average size of interface regions is relatively small compared to the other three regions. The interface regions were found enriched with disease-related nsSNPs ($nsSNPs^{GD}$) (Figure 3.2B) as previously reported [26, 27, 108]. One can observe a two-fold enrichment in interface $nsSNPs^{GD}$ when compared to $nsSNPs^{C}$ (Table 3.2). Moreover, $nsSNPs^{SC}$ were also found to be enriched by 20% (Figure 3.2C) compared to the control

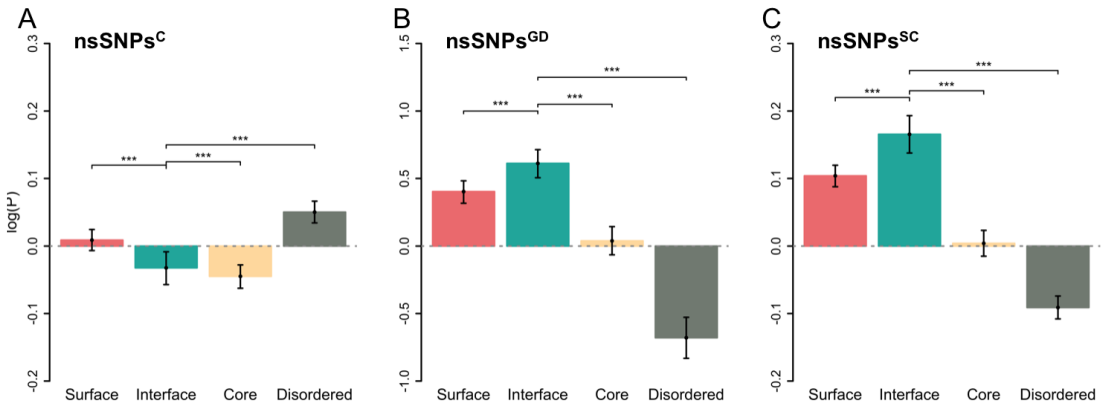

Figure 3.2: The enrichment of nsSNPs at different protein regions. (A) nsSNPs^{C} propensity. There is an enrichment of nsSNPs^{C} at disordered regions. (B) nsSNPs^{GD} Propensity. (C) nsSNPs^{SC} propensity. Both germ-line disease and somatic cancer nsSNPs are enriched at the interface regions of proteins. Propensities are reported in logarithmic scale. Error bars were estimated with bootstrap re-sampling with 10,000 replicates. Stars are drawn to indicate the statistic significance levels ($*p < .05; **p < .01; ***p < .001$) of the comparison between interface and other region nsSNPs from pair-wise Wilcoxon comparison tests.

(Table 3.2).

When comparing ordered (surface, interface and core) and disordered regions of proteins, they are expected to differ in the tolerance to genetic variation, since ordered regions are generally known to have more structural constraints in order to maintain the stability of protein structures. Indeed, both nsSNPs^{GD} and nsSNPs^{SC} showed an overall enrichment at protein ordered regions as previously documented [199], while common nsSNPs were found significantly enriched at disordered regions. On the whole, only a small fraction of nsSNPs$_{disordered}$ was found to be related to disease. This supports our assumption that disordered regions are more tolerant to the occurrence of nsSNPs than ordered protein regions.

In this study, the same enrichment analyses were performed on the nsSNPs' functional specificity and structural constraints. The results section is presented as follows. Firstly, the correlation between structural flexibility and tolerance to genetic variations will be

	Surface	Interface	Core	Disordered
nsSNPs^{GD}	1.48	1.91	1.09	0.48
nsSNPs^{SC}	1.1	1.22	1.05	0.87

Table 3.2: Fold change of disease SNPs dataset. A value less than one indicates a negative fold change (e.g. a value of 0.5 indicates a drop of 50% from the control and is reported as -2 fold change). Both nsSNPs^{GD} and nsSNPs^{SC} were found to be enriched at protein ordered region with respect to the control SNP data, while negative fold changes were observed at disordered regions.

measured by calculating the enrichment of nsSNPs on the secondary structure elements of each protein. The properties of the nsSNPs will also be analysed relative to the transitions of amino acid types and compared between four pre-defined classes of nsSNPs (see Methods) and compared between nsSNP datasets. Secondly, the likelihood of nsSNPs^{GD} and nsSNPs^{SC} affecting protein function by occurring at functionally important positions will be measured by screening nsSNPs that are at or close to protein functional sites and Post-Translational Modification (PTM) sites in the structural 3D space. Finally, we will further investigate the enrichment of nsSNPs^{GD} and nsSNPs^{SC} at interface regions by measuring the frequency of interface co-localisation for disease-related nsSNP pairs involved in the same type of disease. The functional similarity of interaction protein pairs will also be measured. In particular, the pairs that are involved in the same disease will be compared with the pairs related to different diseases and non-disease-related pairs.

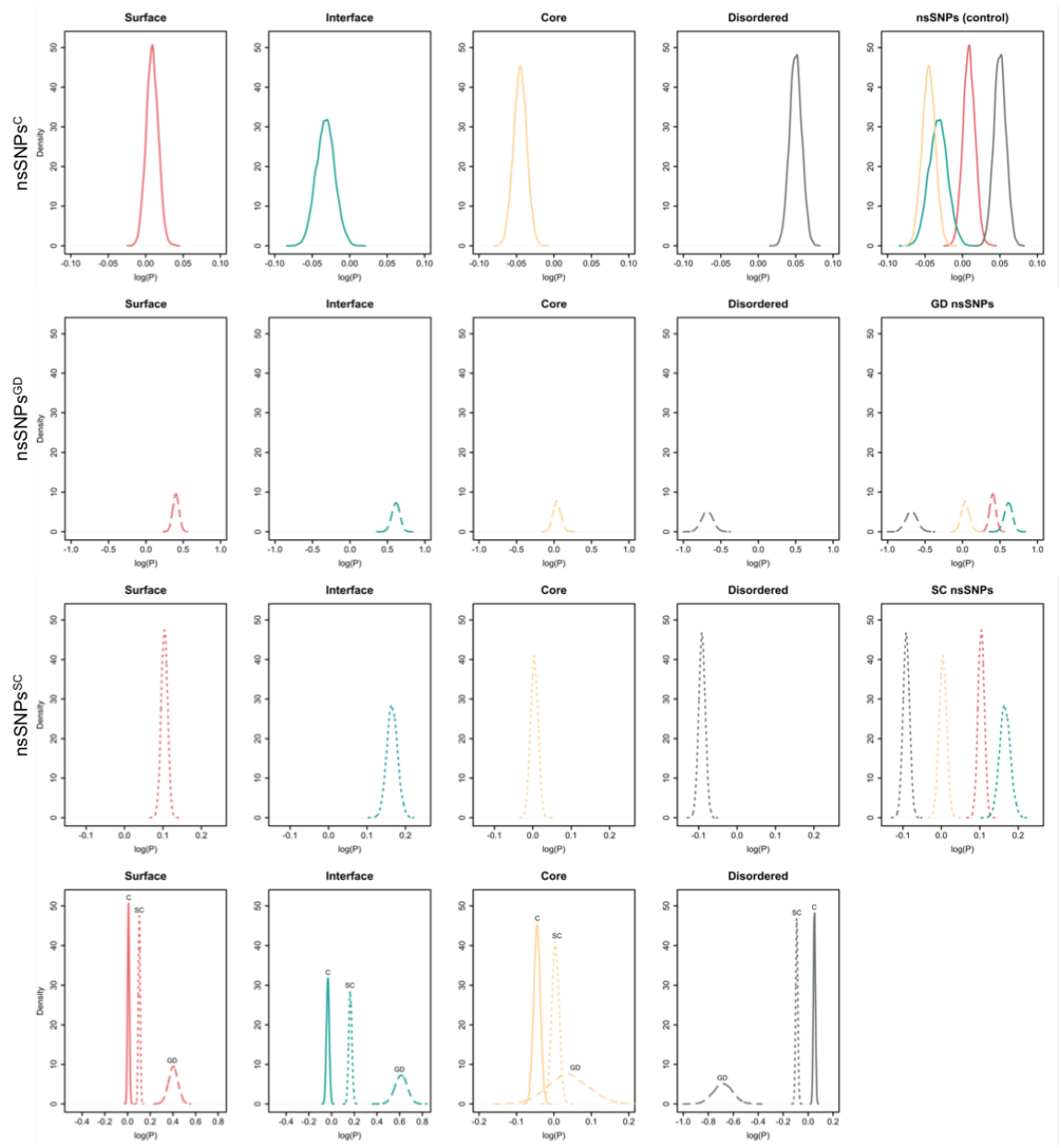

Figure 3.3: Distributions of re-sampled SNP propensities. The first four columns show the distributions of the SNP propensities relative to four protein regions, while the plots in the last column show the propensities at all four protein regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

3.2.2 Structural Features of Disease-related nsSNPs

The relationship between structural constraints and the abundance of disease-related nsSNPs at protein ordered regions (surface, interface, core) was investigated and compared within each nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC} dataset. This analysis was performed by comparing the propensity of nsSNPs to occur at regions with different secondary structure elements. This has commonly used by computational prediction methods as a parameter to evaluate the effect of amino acid substitutions.

Based on the secondary structure definition from DSSP [200], each class at ordered region was further divided into stiff and flexible structural segments (Figure 3.4A). The stiff segments include helix and strand, while loop and turn are classified as flexible segments (see "3.4.5 nsSNPs at Secondary Structure Elements"). A summary of collected nsSNPs in each secondary structure segment, protein region, and nsSNP data group is given in Table A.1.

The stiff segments of proteins are generally considered to be less tolerant to mutations than the flexible segments. In general, nsSNPs^{C} are under-represented in the structured regions of proteins, apart from surface flexible segments which were found to have a higher propensity of nsSNPs^{C} (Figure 3.4B). nsSNPs^{C} were found mostly occurring at disordered regions as reported in the previous section (Figure 3.2).

In comparison, both nsSNPs^{GD} and nsSNPs^{SC} were found to be enriched at ordered regions rather than disordered regions. nsSNPs^{GD} were observed at comparatively higher propensities at the stiff secondary structure elements in the solvent exposed area (including both surface and interface regions of isolated proteins) than the flexible elements (Figure 3.4C, purple bars: nsSNPs^{GD} stiff segments; green bars: nsSNPs^{GD} in flexible segments).

nsSNPs^{GD} were enriched by more than two-fold at the interface stiff segments compared to the control (Table 3.3). These nsSNPs may affect protein functions either by altering the structure at the exposed area or by affecting critical functional sites, which will be further discussed in the next section.

nsSNPs^{GD} at core regions were found to be enriched at flexible segments (Figure 3.4C, green bars and green-coloured secondary structure in Figure3.4A). This may relate to the fact that the structure composition at a protein core region is often critical to maintain the stability of the protein. The occurrences of nsSNPs^{GD} at the flexible segments of the core may have an impact on protein function and stability but are not as critical as nsSNPs^{GD} at the stiff segments.

Whereas, nsSNPs^{SC} were found less enriched at stiff segments for all the considered classes of nsSNPs (Figure 3.5). nsSNPs^{SC} showed less differences in secondary segment preferences at surface regions (Figure 3.5). When compare with nsSNPs^{C}, the nsSNPs^{SC} are enriched by 26% at the interface flexible regions (Table 3.3).

These results may suggest that nsSNPs^{GD} and nsSNPs^{SC} exhibit different physicochemical properties and adopt different mechanisms to affect protein structures and functions. To further assess this hypothesis, this analysis was followed by a more detailed investigation at the residue level by calculating the frequencies of amino acid type changes relative to each class of nsSNPs.

Each amino acid exhibits different characteristics in terms of size, charge, hydrophobicity, polarity and other physicochemical features. An event of residue substitution can be classified as a drastic or moderate change based on the similarity between the two amino acids. A drastic change, such as the change from a positively charged residue to a negatively charged residue, is more likely to have greater impact on the stability of the protein

structure than a moderate change from a positively charged to a polar residue (scheme in Figure 3.6).

The frequency of nsSNPs^{GD} and nsSNPs^{SC} in each drastic and moderate change sub group was calculated and compared with neutral mutation nsSNPs^{C}. In general, nsSNPs^{GD} and nsSNPs^{SC} showed a higher frequency for most of drastic change sub groups compared to nsSNPs^{C} (Figure 3.7 orange bars: nsSNPs^{C}; blue bars: nsSNPs^{GD}; green bars:nsSNPs^{SC}). In particular, nsSNPs^{GD} were found with comparatively higher frequency in drastic changes and lower frequency in moderate changes (Figure 3.7, blue bars).

The frequency of drastic and moderate changes induced by ordered-region nsSNPs (surface, interface, and core) were also calculated relative to the secondary structure elements. Interestingly, drastic substitutions induced by nsSNPs^{GD} showed a higher abundance in stiff structure segments over three classes of nsSNPs^{GD} (Figure 3.8, blue bars). On the other side, each class of nsSNPs^{SC} (Figure 3.8, green bars) showed a pattern similar to neutral mutation nsSNPs^{C} (Figure 3.8, orange bars).

In summary, these results show that nsSNPs^{GD} exhibit features that are consistent and distinguishable from nsSNPs^{C} in terms of structural preferences and trends in physicochemical changes. Whereas, nsSNPs^{SC} have higher enrichment at flexible structural segments and with milder physicochemical changes. However, nsSNP^{SC} dataset contains not only the driver variants but also additional passenger variants. The passenger mutations are found in the cancer genome but do not contribute directly to the development of cancer. The results obtained from nsSNPs^{SC} dataset may be partially ascribed to cancer passenger variants. This could explain the similar feature of the nsSNPs^{SC} to the control.

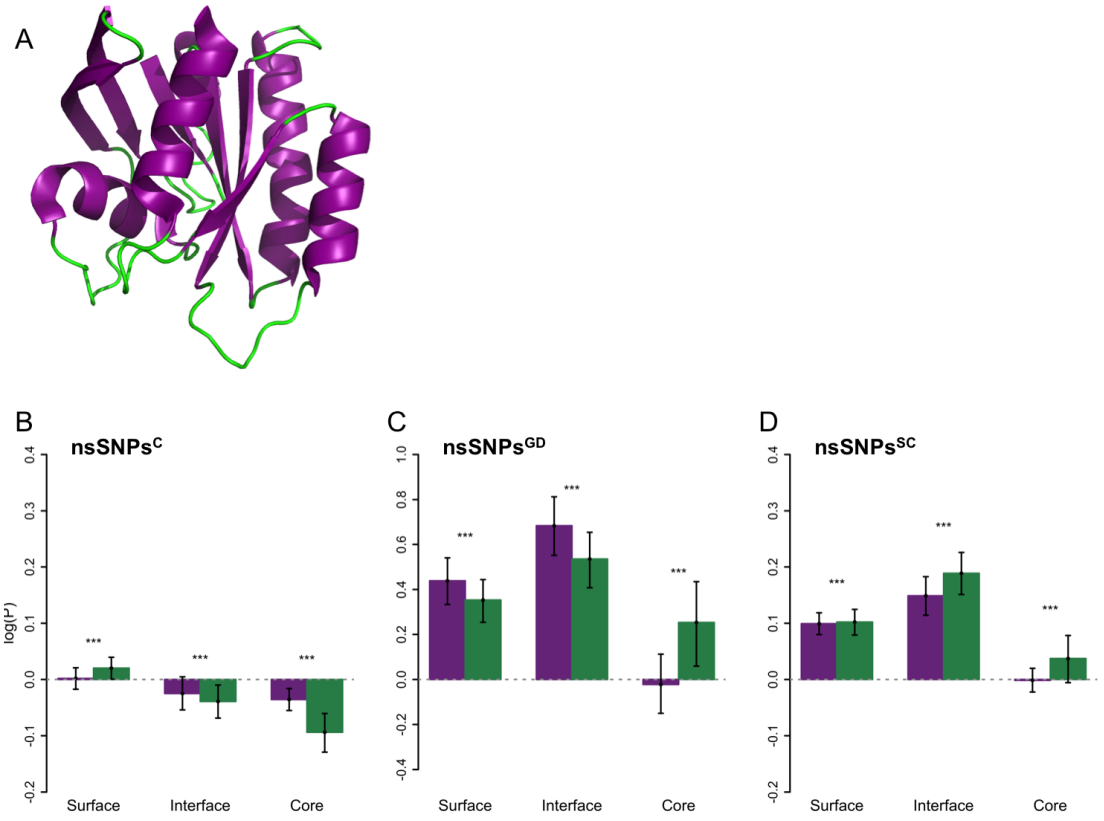

Figure 3.4: nsSNPs associations with protein stiff and flexible structure regions. (A) Stiff (purple) and flexible (green) structure regions. The stiff regions of proteins include helix and strand, while flexible regions include loops and turns. (B-D) The propensities of different nsSNPs (nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}) relative to stiff (purple) and flexible (green) structure regions. The greater the value, the stronger the association with the type of the secondary structure. Propensities are reported in logarithmic scale. Error bars were estimated with bootstrap re-sampling with 10,000 replicates. Stars are drawn to indicate the statistic significance levels ($*p < .05; **p < .01; ***p < .001$) of the comparison between stiff and flexible region nsSNPs from pair-wise Wilcoxon comparison tests. A summary of nsSNPs in different classes of stiff/flexible regions is given in Table A.1.

	Surface^{stiff}	Surface^{flexible}	Interface^{stiff}	Interface^{flexible}	Core^{stiff}	Core^{flexible}
nsSNPs^{GD}	1.55	1.4	2.03	1.78	1.01	1.43
nsSNPs^{SC}	1.1	1.09	1.19	1.26	1.03	1.14

Table 3.3: Fold change of disease SNPs dataset at stiff and flexible structure regions. A value less than one indicates a negative fold change.

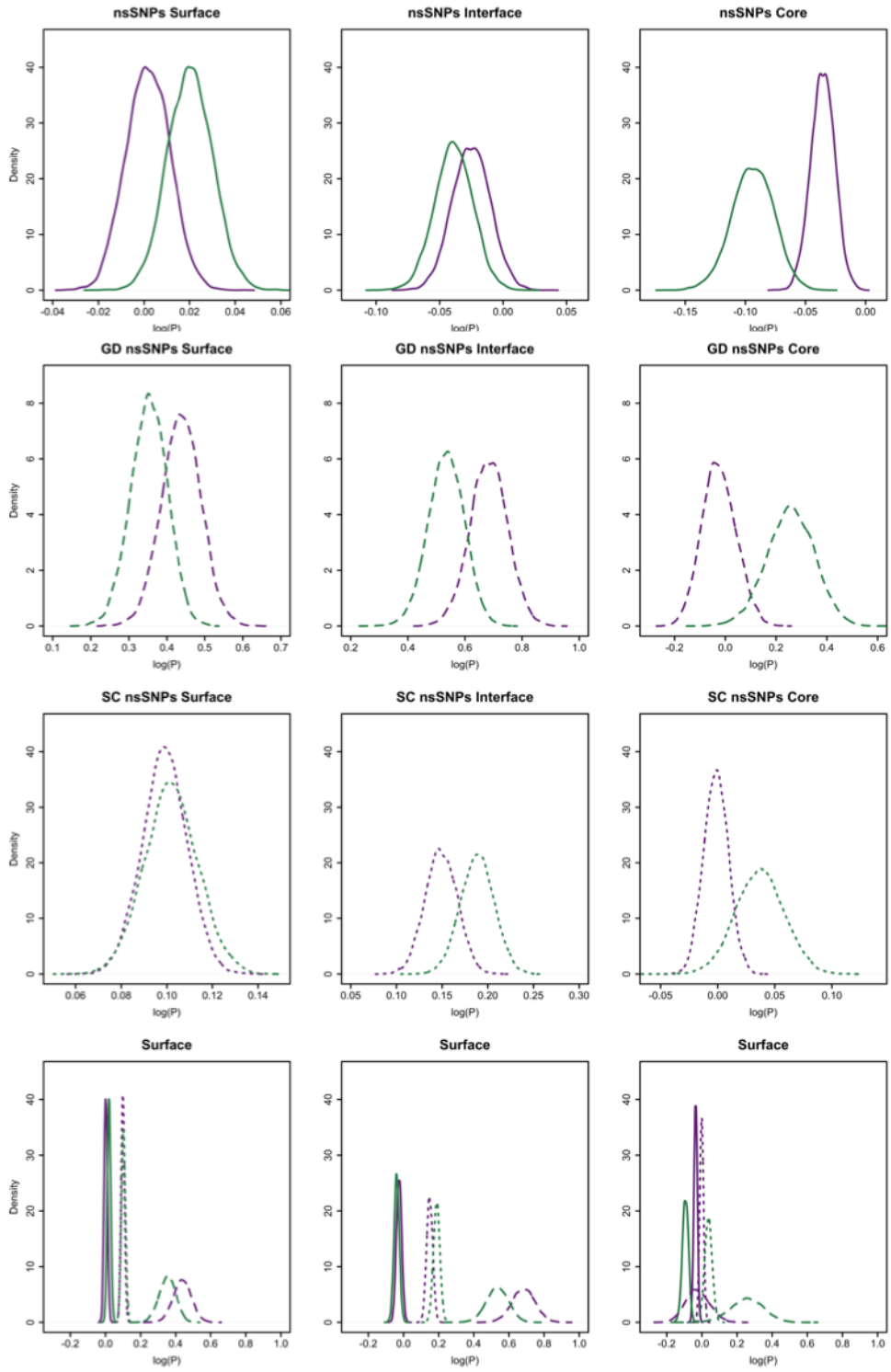

Figure 3.5: Distributions of re-sampled SNP propensities. The columns show the distributions of the SNP propensities relative to protein surface, interface and core regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

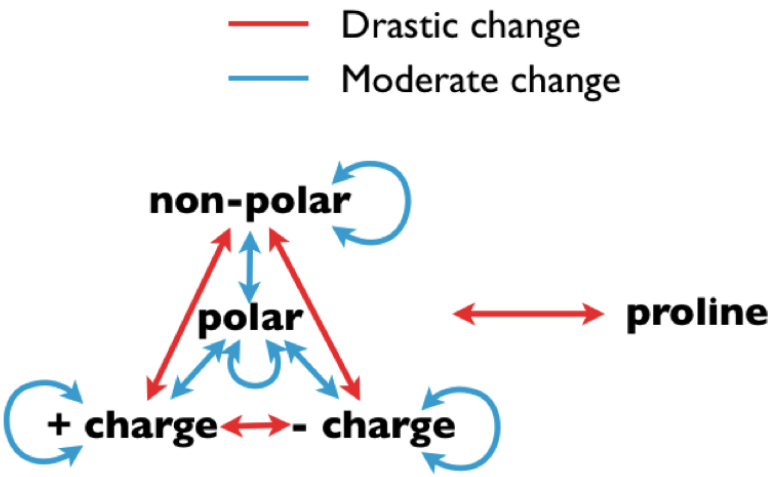

Figure 3.6: Scheme of drastic and moderate amino acid type changes in nsSNPs. Drastic changes (red) include changes between non-polar and negatively charged, non-polar and positively charged, positively charged and negatively charged residues, and any residues change to or from proline. Whereas moderate changes (light blue) include changes between non-polar and polar, polar and negatively charged, polar and positively charged residues, together with changes within each residue class.

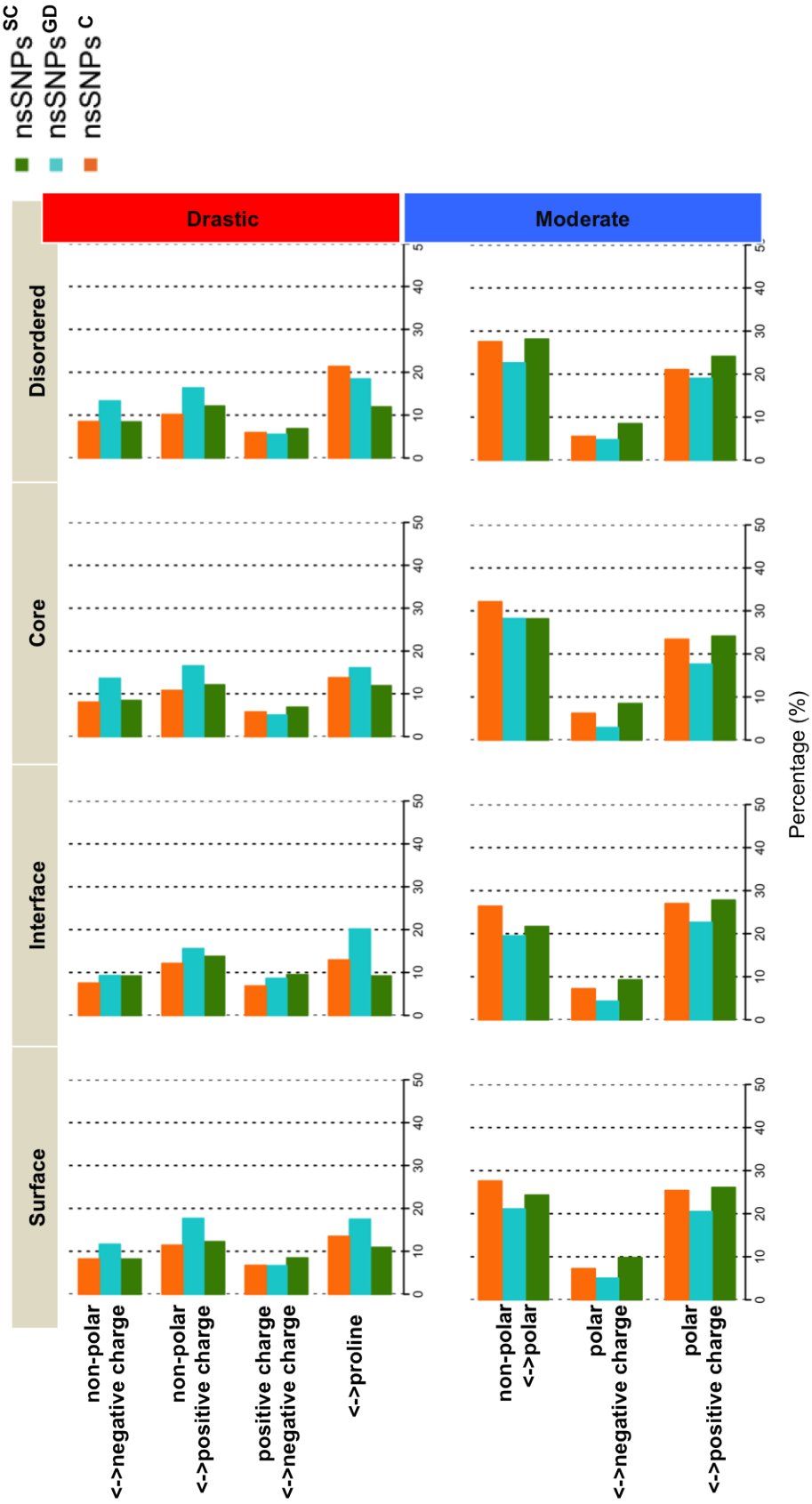

Figure 3.7: nsSNPs drastic and moderate amino acid type changes. Frequency of drastic and moderate amino acid change sub groups calculated for nsSNPs^{C} (orange), nsSNPs^{GD} (blue) and nsSNPs^{SC} (green) at different protein regions Surface, Interface, Core and Disordered. Arrows drawn above nsSNPs^{GD} and nsSNPs^{SC} indicate higher or lower frequencies compared to nsSNPs^{C}.

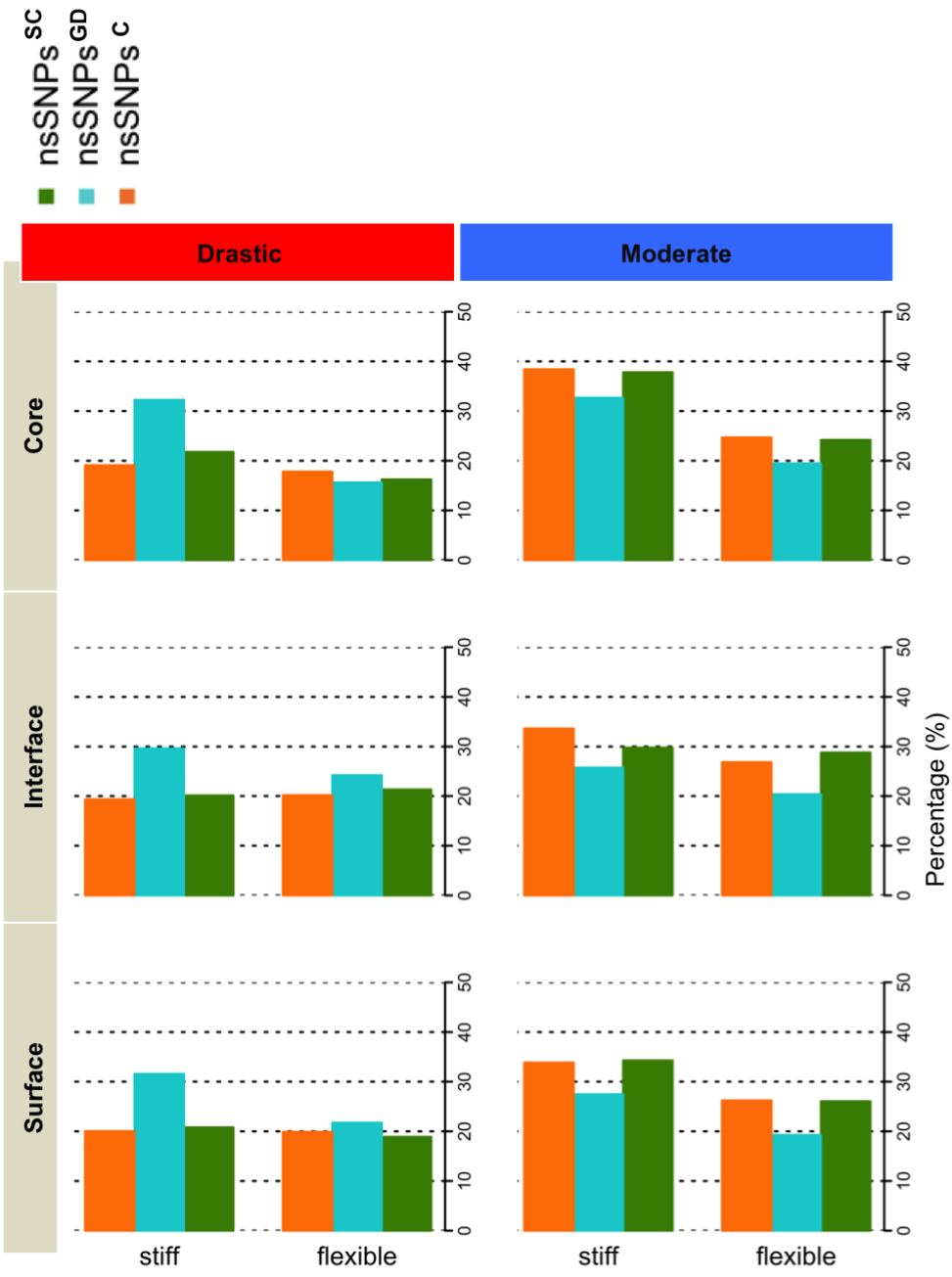

Figure 3.8: nsSNPs drastic and moderate amino acid type changes on secondary structure segments. Frequency of drastic and moderate changes calculated for nsSNPs^{C} (orange), nsSNPs^{GD} (blue) and nsSNPs^{SC} (green) at different protein regions (Surface, Interface, Core and Disordered) and different secondary structure elements. Arrows drawn above nsSNPs^{GD} and nsSNPs^{SC} indicate higher or lower frequencies compared to nsSNPs^{C}. A summary of nsSNPs is given in Table A.2.

3.2.3 Functional Specificity of nsSNPs

Apart from having impact on the stability of protein structures, nsSNPs can also affect protein function through the substitution of functional residues. One example is given by the mutations occurring at protein IDH1 active site. IDH genes encode isocitate dehydrogenase cytoplasmic with the function of catalysing the conversion of isocitrate into α-ketoglutarate [201]. The substitution of an arginine residue at the active site with a different amino acid alters the enzymatic activity of the protein, which consequently elevates the level of R^{-}-2-hydroxyglutarate. This was found to be an important factor to promote the cancer progression of Glioblastoma multiforme [202]. Therefore, to elucidate the molecular features that affect protein function at SNP sites, the occurrences of nsSNPs at functional residues have been examined at both sequence and molecular level. The propensity of nsSNPs to occur at functionally important residue positions have been calculated for each nsSNP class. The annotations of functionally important sites, such as active sites, binding sites, metal binding, PTM sites, were obtained from the UniProt [203] and PTMcode [204] databases. A total of 2,100 and 4,370 proteins in the human 3D PPIN were annotated with functional sites and PTM sites, respectively. The PTM site analysis will be studied independently as they exhibit molecular characteristics different from other functional residues.

A screening of nsSNPs occurring at the exact positions of functionally important residues was first implemented. Only a small number of proteins in the 3D human PPIN were found to have nsSNPs occurring at functional residue positions (see Table A.3 and Table A.4 section A). One likely explanation is that the occurrences of nsSNPs at the functionally important residue positions is more likely to be lethal for the mutant-carrying organism. Therefore, this type of mutations is not fixed in the population.

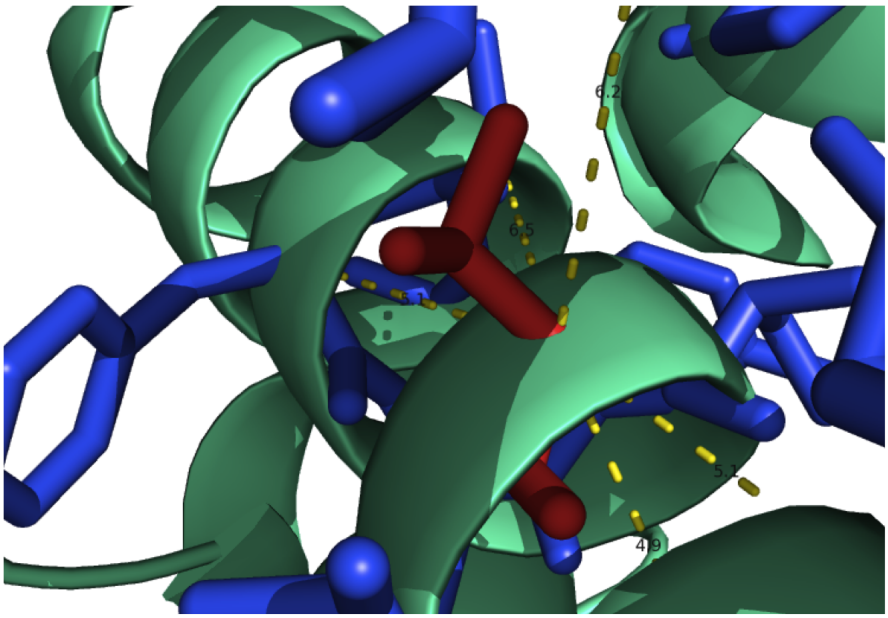

Figure 3.9: Screening SNPs close to functional sites in 3D. Residues spatially close to a given functional site (stick representation coloured in red) were selected using a 8Å threshold on Cα atoms. The residues shown in stick representation and coloured in blue are the residues detected to be close to the considered functional site.

Moreover, the occurrence of nsSNPs close to functionally critical residues can also have an impact on protein function. The V600E mutant of B-RAF, for instance, has been found in many human cancers. The mutation, adjacent to the phosphorylation sites Thr598 and Ser601, affects the cell by stimulating B-RAF kinase activity and consequently increasing ERK signalling [205]. Therefore, spatially proximate residues to functional residues should also be taken into account in the implementation of the screening.

The screening of nsSNPs close to functionally important residues was implemented both at the structural level (Figure 3.9) and at the sequence level. The screening of spatially close residues requires protein structure information and thus was implemented only for the nsSNPs occurring at ordered protein regions, including surface, interface and core. Residues spatially close to a given functional site were selected using a 8Å threshold on Cα atoms (Methods). The sequence-based screening was performed using a screening window of 5 residues. Only the sequence-based screening was used to measure the enrichment

of disordered region nsSNPs that are close to functional residues and to perform the comparison between ordered (surface, interface and core) and disordered regions. The number of nsSNPs that were found to be close to functionally important residues is listed in Table A.3 and Table A.4. Expectedly, the structure-based method detected a higher number of close-to-functional-site nsSNPs (Table A.3C) than the sequence-based method. Moreover, the results from the structure-based screening method showed a slightly different patten for nsSNPs^{GD} group, where the comparative enrichment is more evident for surface nsSNPs^{GD} (Figure 3.10B and E).

The propensities of nsSNPs relative to protein regions were calculated and compared within each nsSNP data group (nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}) Figure 3.10. The propensities calculated for different nsSNP classes in all SNP groups indicates that nsSNPs are generally not favoured to be close to functional residues (propensities are smaller than 1). However, by comparing nsSNPs^{GD} and nsSNPs^{SC} to nsSNPs^{C} data group, both sequence-based and structure-based screening methods detected a comparative enrichment of close-to-functional-site nsSNPs for both disease nsSNP groups. nsSNPs^{GD} showed more than two-fold enrichment for all the ordered protein regions compared to nsSNPs^{C} (Table 3.4). Whereas, the nsSNPs at disordered regions were found to have the lowest frequency in the region close to functional sites in all three nsSNP data groups (Figure 3.10A, B and C and Figure 3.11).

When looking at close-to-PTM-site nsSNPs, all SNP data groups were also found with no preference to be close to PTM sites (propensities are smaller than 1). However, both sequence-based and structure-based screening methods revealed a comparative enrichment for both disease nsSNPs^{GD} and nsSNPs^{SC} data group in comparison to nsSNPs^{C}. In particular, nsSNPs^{GD} and nsSNPs^{SC} at interface regions were found more evidently enriched

(Figure 3.13) by a factor of 2 compared to the control (Table 3.5). Whereas, the disordered regions were found to contain a high number of nsSNPs^{C} but they were less abundant in nsSNPs^{GD} and nsSNPs^{SC} compared with interface regions (Figure 3.14).

To summarise these results, we found that nsSNPs are generally not favoured to be close to functionally important residues. This probably due to the fact that the occurrences of mutations at/close to those functional residues are more likely to lethal. Only by comparing the disease nsSNPs^{GD} and nsSNPs^{SC} data groups to nsSNPs^{C}, a comparative enrichment was found in disease nsSNP groups over all nsSNP classes. In particular, both nsSNPs^{GD} and nsSNPs^{SC} at the solvent-exposed surface area of proteins (surface and interface regions) have a significantly higher tendency to be close to functionally important residues (functional sites and PTMs) than those nsSNPs at the core of proteins. The close-to-PTM-site nsSNPs also showed a larger propensity at disordered regions in general.

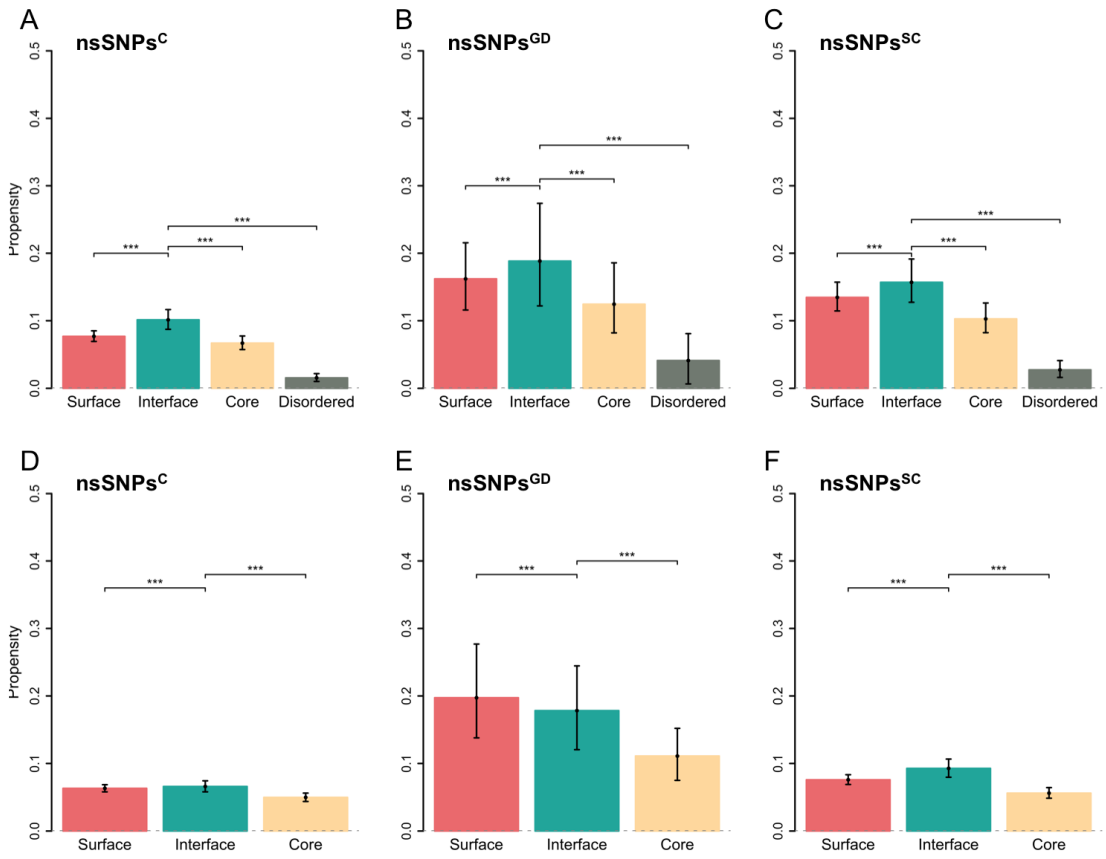

Figure 3.10: nsSNPs close to protein functional sites. Propensities of nsSNPs, which are at or close to functional sites, were calculated over surface (coral), interface (green), core (yellow) and disordered (grey) regions. Error bars were estimated with bootstrap re-sampling with 10,000 replicates. Stars are drawn to indicate the statistic significance levels ($*p < .05; **p < .01; ***p < .001$) of the comparison between interface nsSNPs and other nsSNP classes from pair-wise Wilcoxon comparison tests. (A)-(C) Close-to-functional-site nsSNPs identified by a sequence-based screening method with a window of 5 residues. (D)-(F) Close-to-functional-site nsSNPs identified by structure-based screening method with a 8 Å threshold. (A) and (D) nsSNPs^{C} propensity. (B) and (E) nsSNPs^{GD} propensity. (C) and (F) nsSNPs^{SC} propensity. Interface nsSNPs are comparatively enriched at the segments close to functional sites from both screening methods in the three nsSNP data groups. A summary of detected nsSNPs using both methods is given in Table A.3.

	Surface	Interface	Core
nsSNPs^{GD}	3.11	2.69	2.23
nsSNPs^{SC}	1.2	1.4	1.13

Table 3.4: Fold change of close-to-functional-site SNPs (3D). A value less than one indicates a negative fold change. Both nsSNPs^{GD} and nsSNPs^{SC} show higher enrichment compared with the control SNP data.

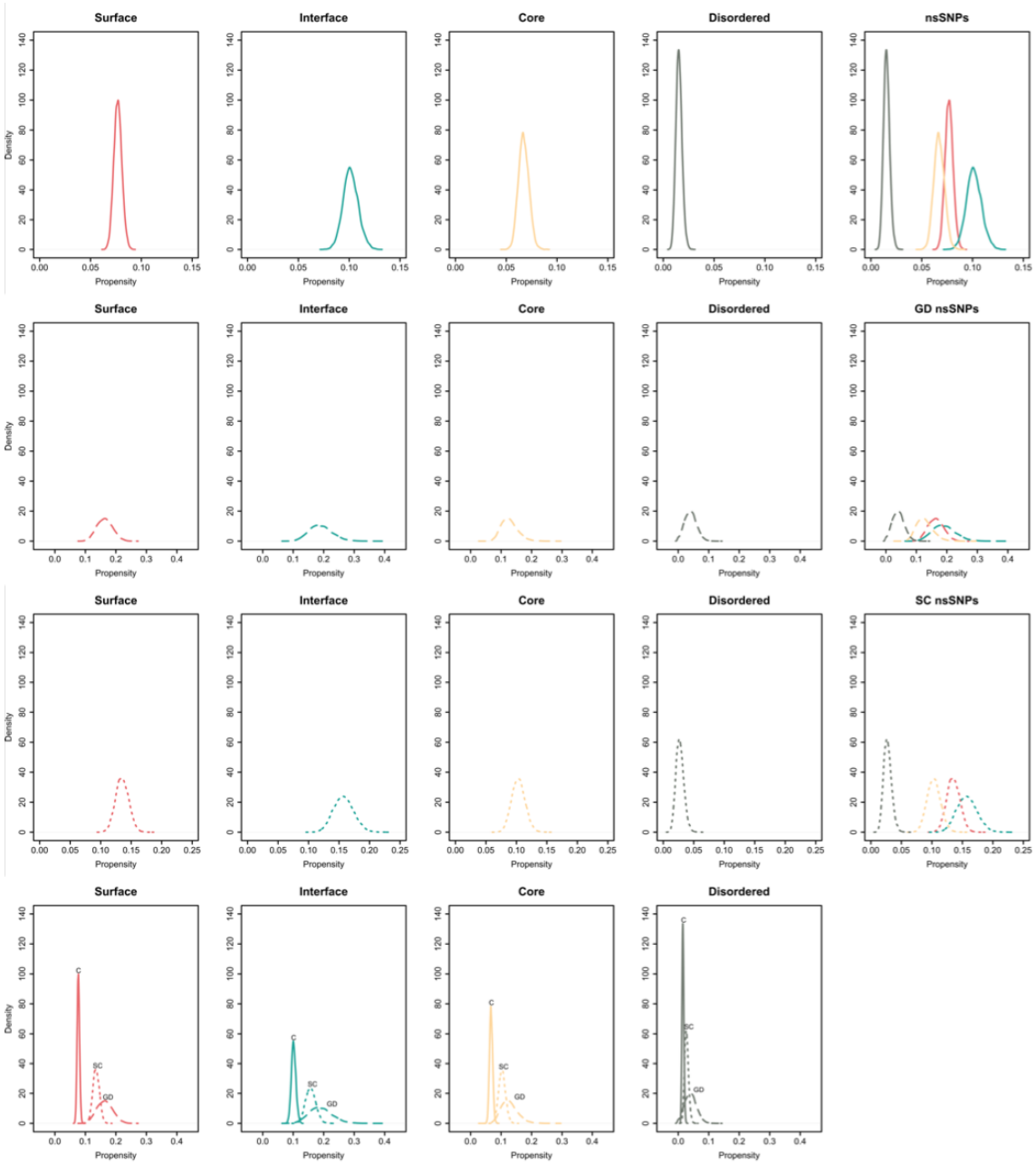

Figure 3.11: Distributions of re-sampled close-to-functional-site SNP propensities. The first four columns show the distributions of the close-to-functional-site SNP propensities relative to four protein regions, while the plots in the last column show the propensities at all four protein regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

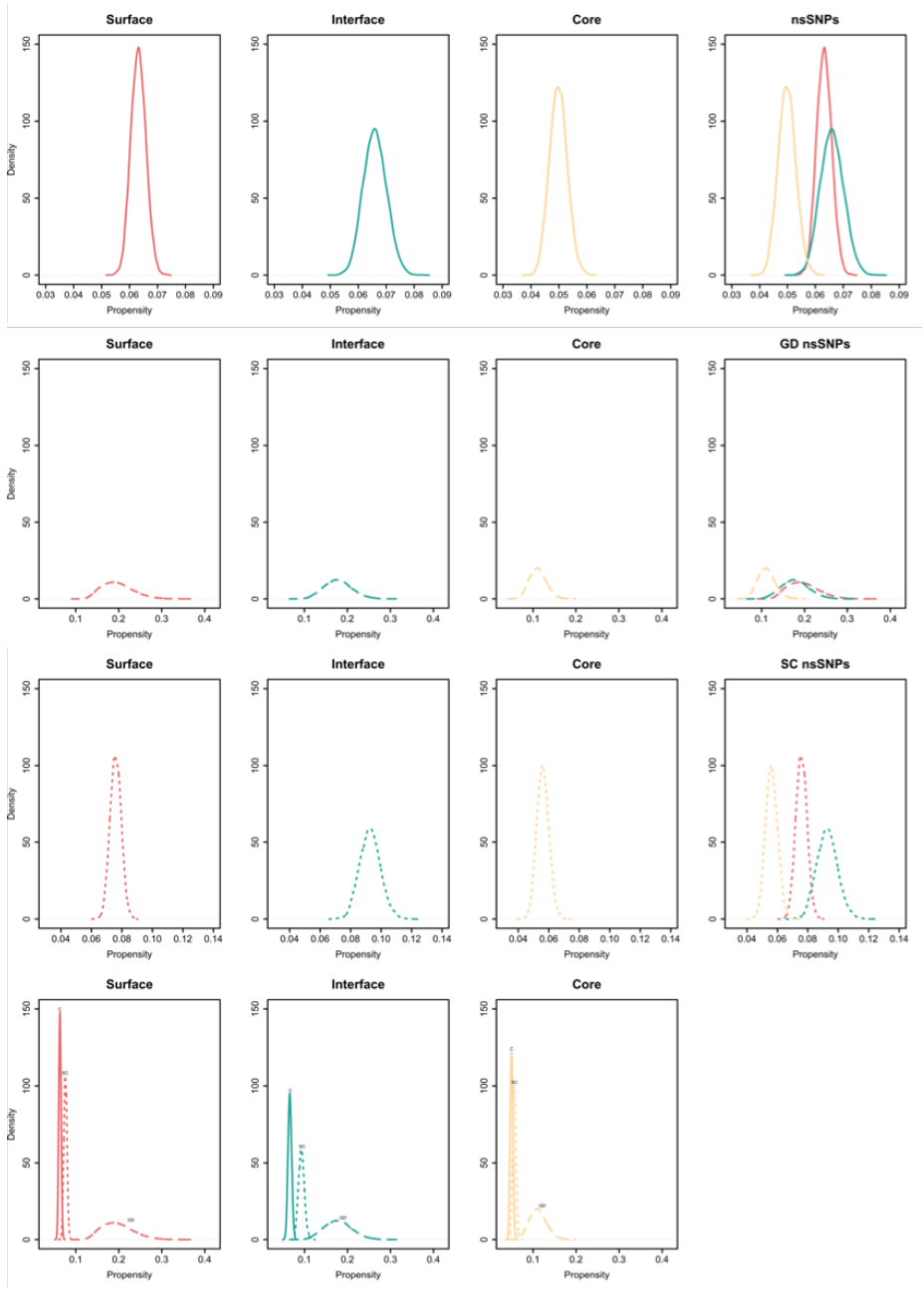

Figure 3.12: Distributions of re-sampled close-to-functional-site SNP propensities (in 3D space). The first three columns show the distributions of the close-to-functional-site SNP propensities relative to protein surface, interface and core regions, while the plots in the last column show the propensities at all three protein regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

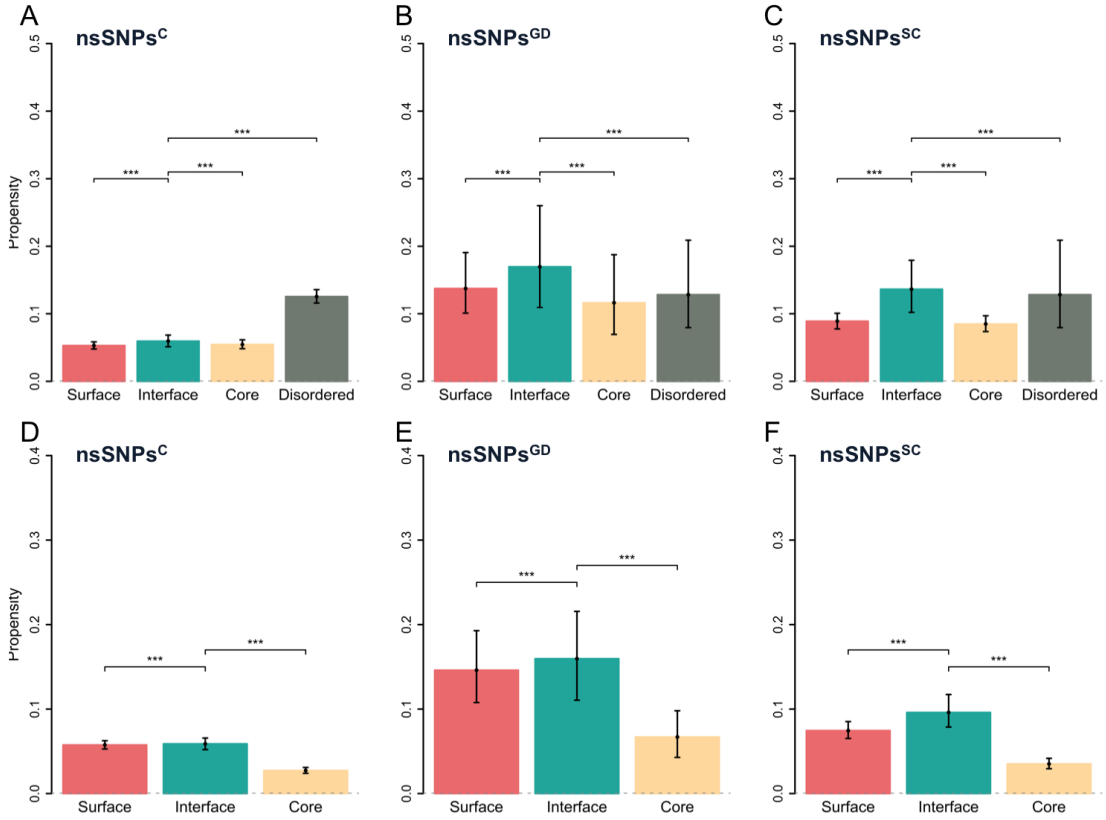

Figure 3.13: Close to protein PTM site nsSNPs. Propensities calculated over Surface (coral), Interface (green), Core (yellow) and Disordered region (grey) nsSNPs that are at or close to PTM sites. Error bars were estimated with bootstrap re-sampling with 10,000 replicates. Stars are drawn to indicate the statistic significance levels ($*p < .05; **p < .01; ***p < .001$) of the comparison between interface nsSNPs and other nsSNP classes from pair-wise Wilcoxon comparison tests. (A)-(C) Close-to-PTM-site nsSNPs identified by sequence-based screening method with a window of 5 residues. (D)-(F) Close-to-PTM-site nsSNPs identified by structure-based screening method with a 8 Å threshold. (A) and (D) nsSNPs^{C} propensity. (B) and (E) nsSNPs^{GD} propensity. (C) and (F) nsSNPs^{SC} propensity. Interface nsSNPs are comparatively enriched at the segments close to PTM sites from both screening methods in the three nsSNP data group. Disordered regions were also found with a larger propensity in close-to-PTM-site nsSNPs. A summary of detected nsSNPs using both methods is given in Table A.4.

	Surface	Interface	Core
nsSNPs^{GD}	2.53	2.71	2.46
nsSNPs^{SC}	1.29	1.63	1.29

Table 3.5: Fold change of close-to-PTM-site SNPs (3D). A value less than one indicates a negative fold change. Both nsSNPs^{GD} and nsSNPs^{SC} were found with higher enrichment compared with the control SNP data.

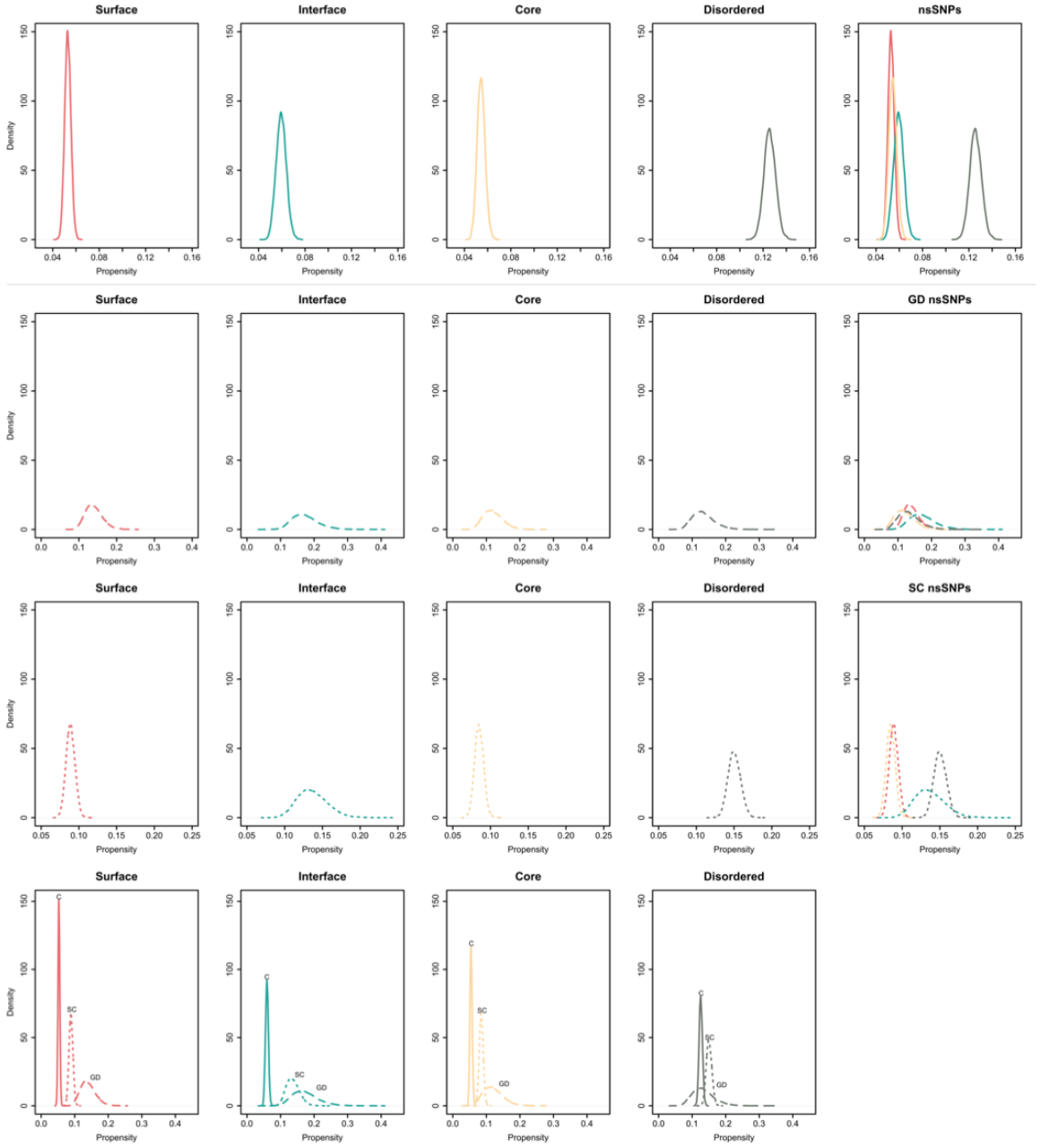

Figure 3.14: Distributions of re-sampled close-to-PTM-site SNP propensities. The first four columns show the distributions of the close-to-PTM-site SNP propensities relative to four protein regions, while the plots in the last column show the propensities at all four protein regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

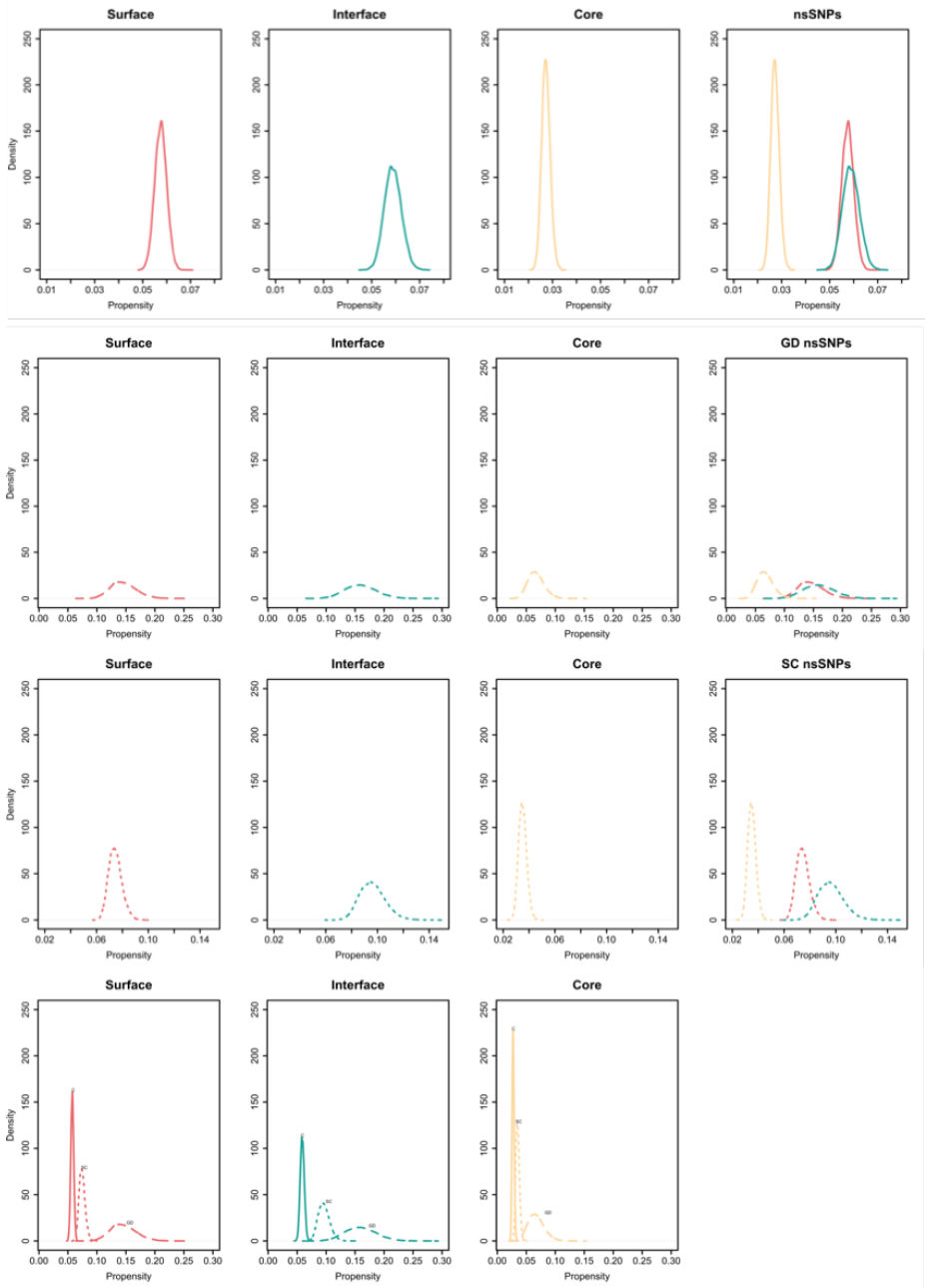

Figure 3.15: Distributions of re-sampled close-to-PTM-site SNP propensities (in 3D space). The first three columns show the distributions of the close-to-PTM-site SNP propensities relative to protein surface, interface and core regions, while the plots in the last column show the propensities at all three protein regions. The top three rows are the propensities of nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}, respectively, at different protein regions. The plots in the last row show the distributions of the three SNP datasets at different protein regions.

3.2.4 Co-localised Disease-related nsSNPs

A previous study [109] elucidated interface co-localised mutations and demonstrated that recessive mutations co-localised on interface of a protein interaction pair have the tendency to cause the same disease. This observation can be explained by the fact that interacting proteins often perform similar biological functions and are involved in the same biological process. However, it is not known why the tendency to interface co-localisation of mutations related to the same disease is observed only for recessive and not for the dominant mutations.

In the study presented here, a different approach was applied to study interface co-localised nsSNPs. First, disease-related nsSNPs (nsSNPs^{GD}) were classified with the associated disease types obtained from Goh *et al.* [10] (see Methods). nsSNPs^{SC} were classified by the primary cancer tumour type. The list of disease types and the number of nsSNPs corresponding to each each disease type are shown in Table A.5 and A.6.

Second, the interface co-localised nsSNPs, which share the interfaces with other disease-related nsSNPs, were further divided into three classes: inter-interface specific (nsSNPs$_{S_Inter}$), intra-interface specific (nsSNPs$_{S_Intra}$) and others (nsSNPs$_{Multi}$). A nsSNP$_{S_Inter}$ is defined when the nsSNP is found to cause the same type of disease as another nsSNP on the opposing side of the interface as shown in Figure 3.16A. A nsSNP$_{S_Intra}$ is a SNP that is found to cause the same type of disease as another nsSNP which is co-localised at the same interface as shown in Figure 3.16B. The interface disease-related nsSNPs which do not fit into these two definitions are classified as nsSNPs$_{Multi}$. The implementation of screening and classifying interface nsSNP is presented in Methods "3.4.11 Classification of Interface nsSNP". The propensities of the interface classes of nsSNPs^{GD} and nsSNPs^{SC} datasets

were calculated and compared using statistic assessment (see Methods).

The comparison of different classes of interface nsSNPs^{GD} showed that the nsSNPs$_{S_Intra}$ were significantly more abundant among interface nsSNPs^{GD} (Figure 3.16C). This may suggest that the causes of germ-line diseases are more likely related to the structural or functional alteration of a single protein or a group of proteins. It may take evolutionary times to accept two substitutions, possibly aided by other compensatory substitutions that were not monitored here. The co-localised nsSNPs^{GD} of a specific protein do not necessarily occur at the same time but they can disrupt the protein function through the same mechanism by altering the same binding interface. The nsSNPs^{SC} group showed instead a reduced propensity for either inter- or intra-interface co-localisation (Figure 3.16D)

A possible explanation of interface co-localisation for nsSNPs^{GD} is that disease types are correlated with protein functions. To assess this hypothesis, the functional similarity between the affected interaction protein pairs was measured. Each protein in the 3D PPIN was annotated with disease types extracted from their relative disease-related nsSNPs. According to the disease types of protein pairs, the interaction pairs were divided into three classes, including the PPIs that are involved in the same disease (PPIs$_{disease}$), involved in different diseases (PPIs$_{diff}$) and that are non-disease related (PPIs$_{non}$). Each protein was annotated with Gene Ontology (GO) term annotations (biological process). The functional similarity of the interacting protein pairs over these three classes of PPIs was measured using the Total Ancestry Similarity (TAS) method and compared (Methods). The association between the resulting disease types and the protein functions was measured by the tendency of nsSNP co-localisation and the functional similarity of the interaction proteins. The lower the TAS score, the higher the similarity between a protein pair.

Our results showed that $\text{PPIs}^{GD}_{disease}$ got significantly lower TAS score than PPIs^{GD}_{diff} and non-disease PPIs (Figure 3.17A). This indicates that $\text{PPIs}^{GD}_{disease}$ have high functional similarity between interaction protein pair, while PPIs^{GD}_{diff} showed the lowest functional similarity. Based on this observation, we may argue that for germ-line diseases, protein functions are an important factor to relate a nsSNP to a specific disease. Moreover, the results in the previous sections indeed showed that nsSNPs^{GD} affect protein functions dominantly by their direct impact on functionally important sites. Those nsSNPs are mutations that interfere with protein functions but are not lethal and transmitted through generations.

Moreover, to support our argument that protein function play a crucial role in determining the associated disease types of nsSNPs, protein domains were also annotated with domain functions using SCOP superfamily functional annotations (Methods). By measuring the frequency of nsSNPs^{GD} related to SCOP [206] domain functions, the associated disease types were found to be related to the functions of the affected proteins (Figure 3.18). For example, a large number of nsSNPs occurring at domains with the SCOP domain function "Metabolism" are associated with Haematological disease.

On the other hand, nsSNP^{SC} proteins showed a very different pattern, where $\text{PPIs}^{SC}_{disease}$ were found to have the lowest functional similarity compared to the rest (Figure 3.17B). This may highlight the fundamental differences between germ-line diseases and somatic cancers, in that the progress of cancer is often suggested to involve multiple mutation events [207].

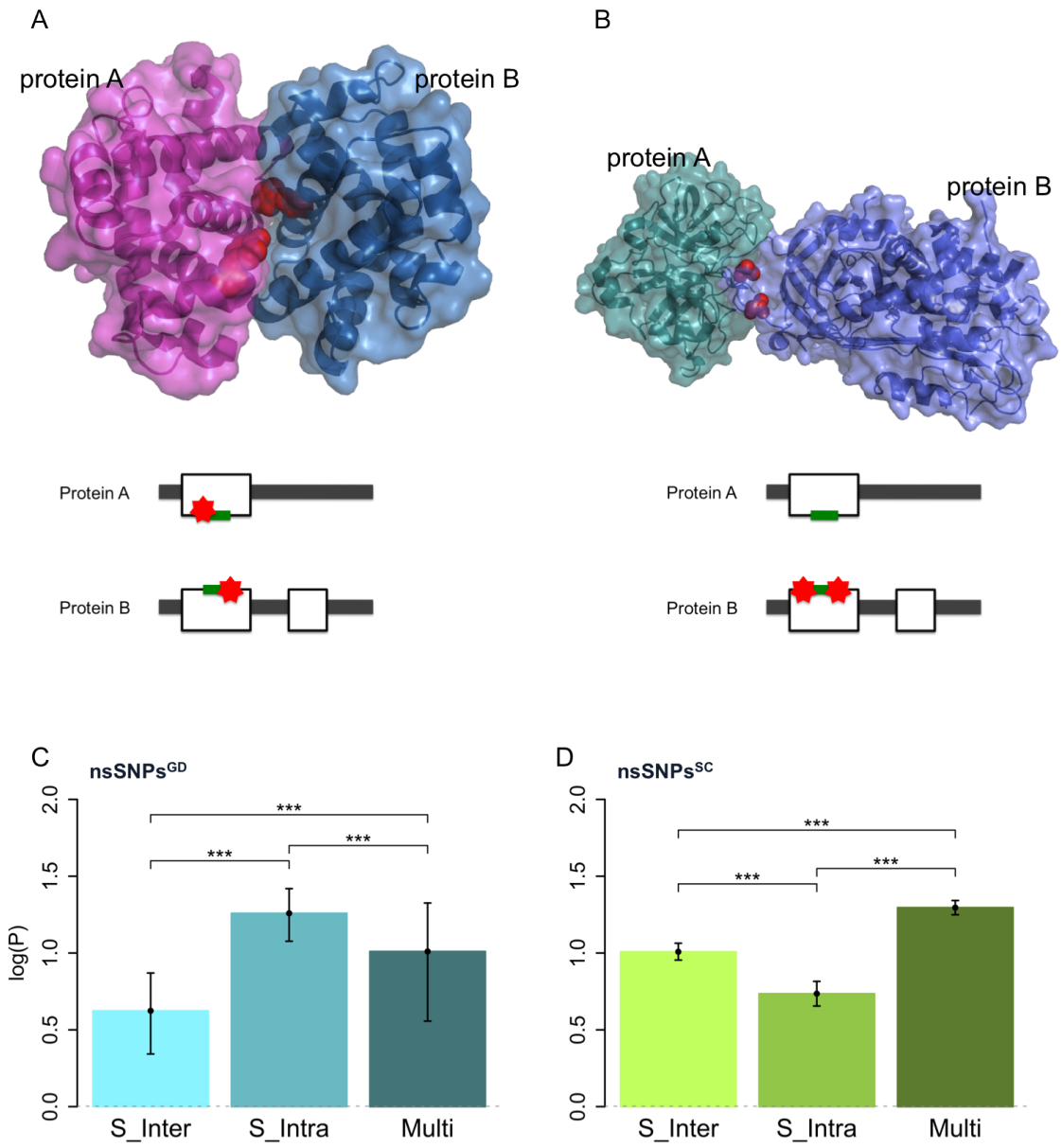

Figure 3.16: Interface co-localised disease-related nsSNPs. (A) Inter-interface specific nsSNPs. Interface nsSNP pairs (residues shown in sphere representation and coloured in red) locate on opposing side of interface of interacting proteins A and B. (B) Intra-interface specific nsSNPs. Interface nsSNPs pairs locate on the same side of the interface. Interface nsSNPs were divided as S_Inter, S_Intra and Multi. S_Inter and S_Intra indicate the nsSNPs shown in (A) and (B), respectively. Interface nsSNPs that do not pair with other nsSNPs related to the same diseases are defined as Multi. (C) and (D) are the propensity of each type of interface nsSNPs^{GD} and nsSNPs^{SC} respectively. Propensities are reported in logarithmic scale. Error bars were estimated with bootstrap re-sampling with 10,000 replicates. Stars are drawn to indicate the statistic significance levels ($*p < .05; * * p < .01; * * *p < .001$) of the comparison between three classes of interface nsSNPs from pair-wise Wilcoxon comparison tests.

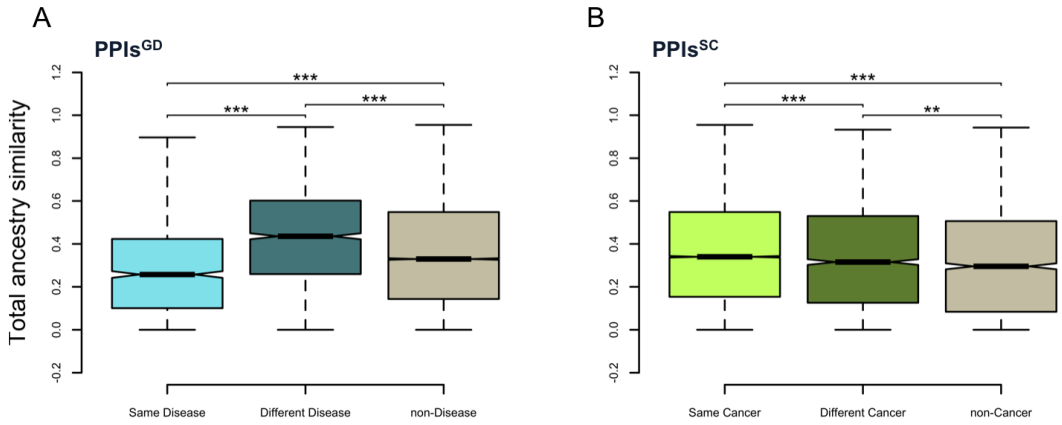

Figure 3.17: Functional similarity of interaction protein pairs. $\text{PPIs}_{disease}$ indicates PPIs that are involved in the same disease and labelled as "Same Disease". PPIs_{diff} indicates PPIs that are involved in the different disease and labelled as "Different Disease". PPIs_{non} indicates non-disease related protein pairs and labelled as "non-Disease". (A) The functional similarity of germ-line disease protein interaction pairs. (B) The functional similarity of somatic cancer protein interaction pairs. Stars are drawn to indicate the statistic significance levels ($*p < .05; * * p < .01; * * *p < .001$) of the comparison between different interface nsSNPs groups from pair-wise Wilcoxon comparison tests.

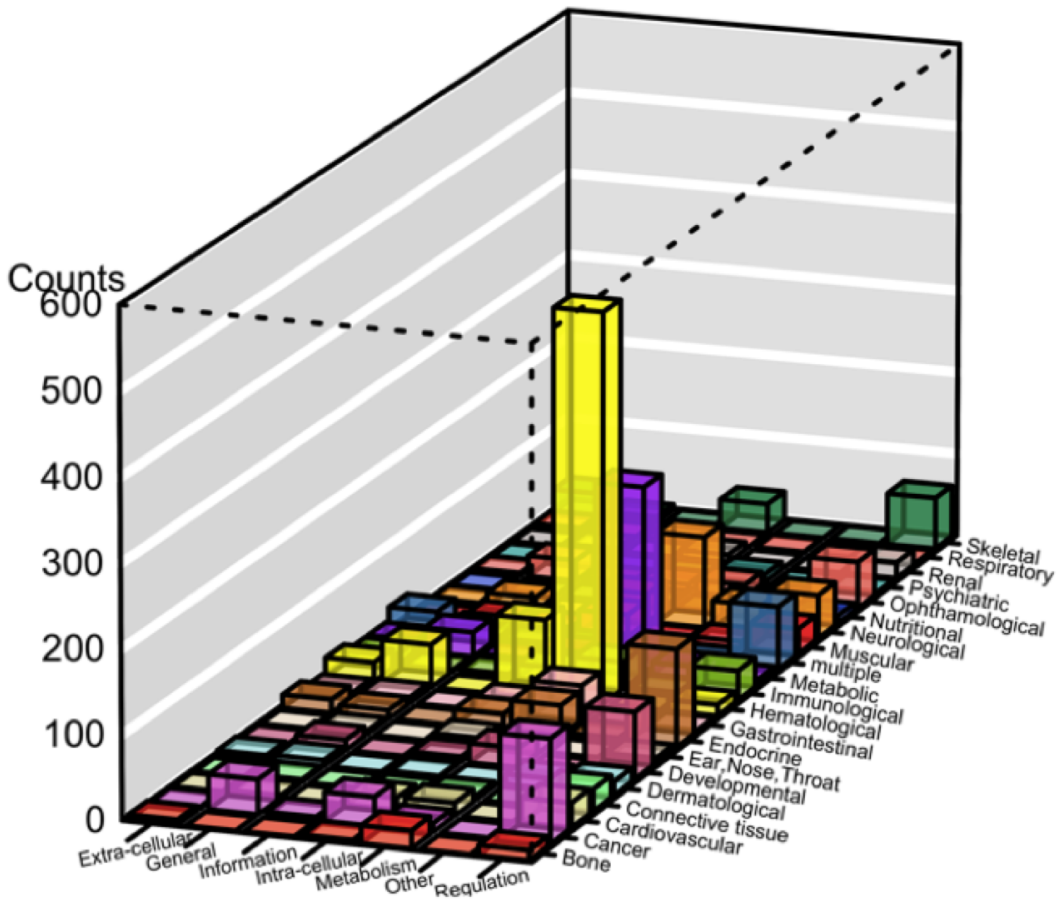

Figure 3.18: Numbers of nsSNPs^{GD} relative to domain functions and the types of diseases. The x axis indicates the SCOP functional categories of domains [206]. nsSNPs^{GD} were assigned to the functional categories of their occurring domains. The nsSNPs^{GD} in each category were further divided by their associated disease types, which are shown in the y axis. The disease categories were obtained from the annotation of Goh *et al.*[10]. The z axis indicates the count of nsSNPs^{GD} in each functional category and the disease type.

3.2.5 Prediction of nsSNP Impact

Current prediction methods, such as PolyPhen2, often use both sequence and structure information to predict the impact of SNPs on protein function. PolyPhen2 was used to predict the impact of nsSNPs in three SNP datasets: nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}. Apart from disordered region nsSNPs, PolyPhen2 performs with fairly good accuracy in germ-line disease SNP predictions, where interface nsSNPs^{GD} got the highest score.

As expected, a larger number of nsSNPs^{SC} were predicted to be "neutral" and in many cases PolyPhen2 was unable to predict the effect of nsSNPs^{SC}. The reasons are that, firstly, the prediction tool was designed for and trained with germ-line mutation SNPs. Moreover, somatic cancer SNPs exhibit structural and functional properties that are rather different from germ-line disease SNPs, as the results in the previous sections have shown.

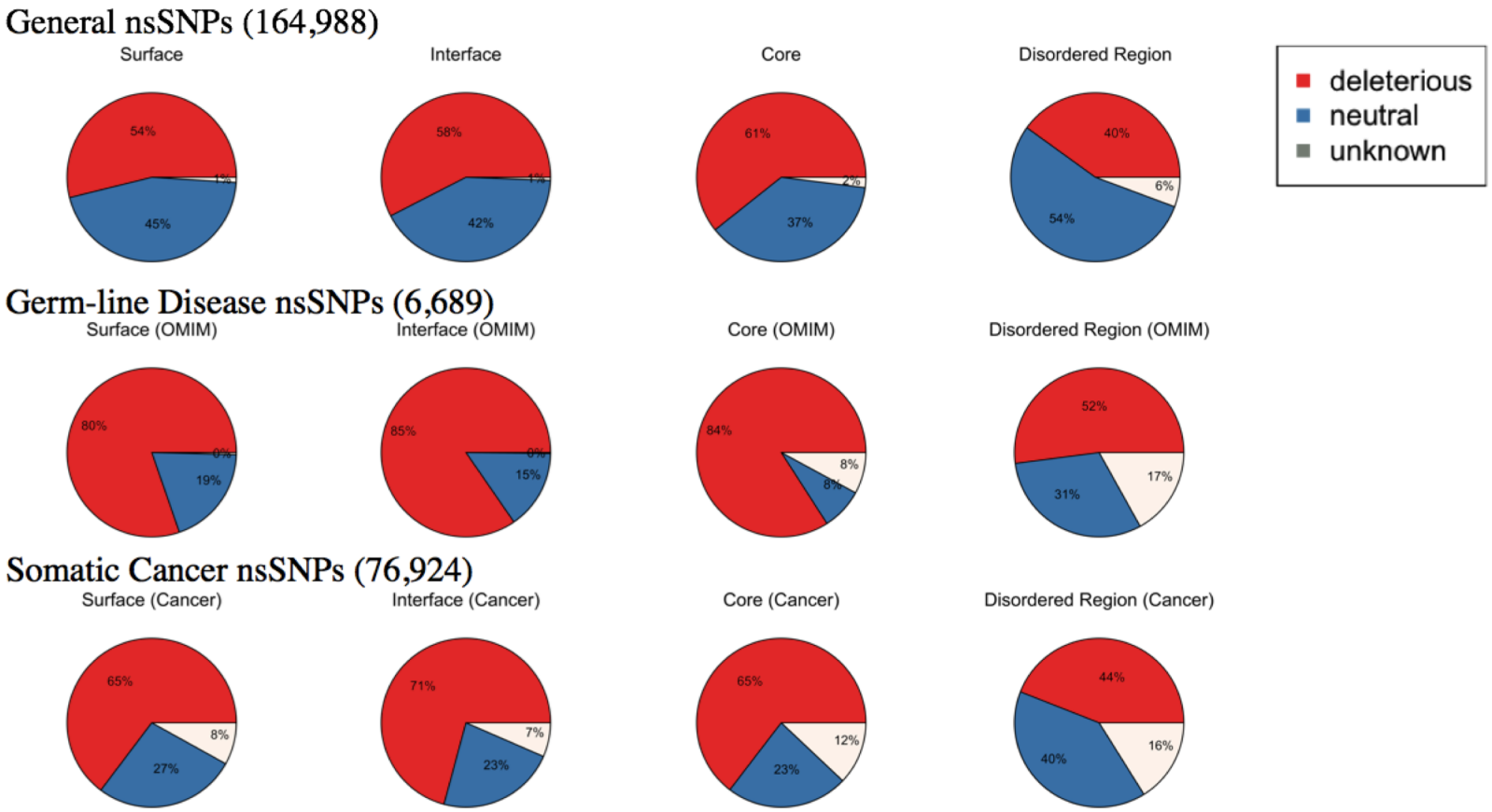

Figure 3.19: Prediction of nsSNPs impact from Polyphen2.

3.3 Conclusions

In this study we looked at the structural and functional features of germ-line disease nsSNPs and somatic cancer nsSNPs to assess their roles on protein function. Our results suggest different behaviours of these two types of disease-related nsSNPs. The nsSNPs^{SC} were found to be enriched at flexible secondary structural elements of proteins and they preferentially locate close to functionally important sites. On the other hand germ-line disease nsSNPs are preferentially located at structurally stiff region and affect protein functions when drastic amino acid changes are fixed in the genome. These results sketch the different constraints acting on these two types of disease-related nsSNPs and the associated disease-causing likelihoods. Moreover, germ-line diseases were found to be highly correlated to the functions of nsSNP-affected proteins. Overall, somatic cancer SNPs were found to exhibit fundamentally different characteristics from germ-line disease SNPs in terms of physicochemical properties and functional specificity. Our study has demonstrated an effective method to characterise disease-related nsSNPs by integrating structural and functional information.

3.4 Materials and Methods

3.4.1 The human 3D Protein-Protein Interaction Network

The human PPI dataset was generated using six publicly available databases, including BioGRID [60], DIP [61], HPRD [59], IntAct [62], MINT [63] and STRING [64]. The dataset from Havugimana *et al.* [67], which contains novel PPIs derived from AP-MS experiments, was also obtained and integrated with the other six PPI datasets. The UniProt [203] accession identifier was used as the reference identifier of the proteins in the network. The

mapping between UniProt accession identifiers and other protein identifiers was based on the mapping provided by UniProt, which was downloaded from the UniProt FTP repository. Having removed duplicated interactions, our human PPI dataset contains 18,399 proteins with 1,878,098 interactions. Each protein was then annotated with the relative Pfam[82] domains. Solved or homologous structures of protein domains were identified running PSI-BLAST v.2.2.26+ [208] against the PDB [79]. Homologous structures were selected when they have more than 80% coverage of a protein domain with over 30%sequence identity. The interaction protein pairs retained in the network had to fulfil the following criteria:

1. An interaction protein pair is constituted of both human proteins.
2. Both proteins have known UniProt accession identifiers.
3. The interaction is non-redundant in the network.
4. A protein pair has at least one solved or homologous structures in the same PDB file (biological unit). To determine the interaction interface region, only the structure of a interaction protein pair with highest sequence identity and coverage was selected.

3.4.2 Interface Identification

For each interaction protein pair, interface residues were defined by analysing the PDB binary complexes using the POPSCOMP [102] tool, which calculates residue solvent accessibility and identifies the exposed surface area of molecules. Protein binding interface residues were determined by subtracting the solvent accessible surface area (SASA) of the protein complex from that of the two isolated proteins:

$$\Delta SASA(A:B)_{interface} = SASA_A + SASA_B - SASA_{A:B} \quad (3.1)$$

where SASA(A:B)$_{interface}$ are the residues located at the binding interfaces of protein A and B complex.

An interaction protein pair is selected for 3D PPIN construction if the interface region of the protein complex is identified. Structured and unstructured regions were identified for each protein in the human 3D PPIN (Figure 3.1). The former includes surface, interface and core regions, while the unstructured regions refer to disordered region that are outsides of Pfam domain boundaries and are predicted as disordered by DISOPRED [176]. Details of determination of surface, interface, core and disordered regions are given in Chapter 4.

3.4.3 Data Collection and Analysis of SNPs

Human gene variation datasets were downloaded from databases, including dbSNP (build 135) [123], OMIM (November 2013 downloaded) [124] and COSMIC (v66_250713 release) [125]. Only single point missense mutations, which change amino acids on protein sequences, were selected for this study. The initial number of nsSNPs in each dataset is listed in Table 3.6 Section *a*. The dataset from dbSNP was used as the control dataset and is referred to as common nsSNPs (nsSNPs^C), while the datasets obtained from OMIM and COSMIC databases represent germ-line disease nsSNPs and somatic cancer nsSNPs, respectively.

This study uses the UniProt accession identifier database as a protein reference. The initial dbSNP dataset contains 800,332 missense mutations referenced with RefSeq protein

	dbSNP (nsSNPs^{C})	OMIM (nsSNPs^{GD})	COSMIC (nsSNPs^{SC})
a. Initial numbers collected from databases :			
nsSNPs	800,332	24,932	1,023,837
b. After mapping to UniProt sequences:			
nsSNPs	759,431	19,002	912,816
c. After filtering data with criteria (as described in the text):			
nsSNPs	480,762	12,761	202,719
proteins	17,744	2,031	17,741
d. After mapping on human 3D PPIN :			
nsSNPs	165,288	6,698	76,924
proteins	8,029	1,115	7,746

Table 3.6: Numbers of nsSNPs collected from databases. The number of nsSNPs in each dataset in four stages of filtering procedure is listed. The details in filtering nsSNP datasets are given in "3.4.3 Data Collection and Analysis of SNPs".

identifiers. NCBI protein database assigns each isoform of a protein with an unique RefSeq identifier, while UniProt assigns each protein with one accession identifier. This could lead to duplicate counts in SNP numbers and mismatch nsSNP positions when mapping from RefSeq to UniProt. In some cases, sequences of the same protein in NCBI and UniProt differ slightly. A nsSNP in the control dataset, which was found at a inconsistently documented residue position, was retained in the dataset. After mapping protein identifiers from RefSeq to UniProt and filtering out mismatched protein residues due to different protein isoforms, the dataset contains 759,431 nsSNPs. The dataset was further filtered with the following criteria:

- nsSNP^{C} records do not exist in either nsSNP^{GD} or nsSNP^{SC} dataset.
- Only the nsSNPs^{C} which occur in the proteins of human PPIN are selected.

The resulting control dataset (nsSNPs^{C}) contains 480,762 nsSNPs in 17,744 proteins.

Germ-line disease nsSNPs which mapped on the residues that differed between the two databases were filtered out from nsSNP^{GD} dataset. An example is given in Figure 3.20. The sequence of ADRB1 (UniProt accession Identifier P08588 RefSeq identifier NP_000675) is documented differently at residue position 389, with an Arginine in UniProt database

and a Glycine in NCBI dataset. This residue has been reported to be related to cardiovascular disease (p.R389G; rs1801253) [209]. Therefore, the checking of residue indexes and amino acids between the sequences of RefSeq and UniProt is required. A total of 71 nsSNPs^{GD} were found at inconsistently documented residue positions. Having filtered out those nsSNPs and the nsSNPs that mapped on mismatched isoform sequences, 19,002 nsSNPs remained in the nsSNP^{GD} dataset. The dataset was then further filtered. Only the nsSNPs^{GD} which occur in the proteins of human PPIN were selected. At this stage, nsSNP^{GD} dataset contained 12,761 nsSNPs in 2,031 proteins.

The third nsSNP dataset contains somatic cancer missense variants obtained from the COSMIC database, which documents somatic cancer mutations and related information curated from the primary literature. The initial number of mutations collected from the database is 1,524,611 cancer mutations, among which 1,023,837 are missense variants. The nsSNPs are referenced with gene names and HGNC idnetifiers (HUGO Gene Nomenclature Committee) [210]. Having mapped gene names with UniProt identifiers, 912,816 nsSNPs remained in the nsSNP^{SC} dataset. The dataset was further filtered with the following criteria:

- A nsSNP is selected if it is annotated as "Substitution - Missense" and "Confirmed somatic variant".
- Only the nsSNPs^{SC} which occur in the proteins of human PPIN are selected.

The resulting somatic cancer mutation dataset (nsSNPs^{SC}) contains 202,719 nsSNPs in 17,741 proteins.

Figure 3.20: An example of differently documented protein sequences between NCBI and UniProt databases. The residue at position 389 in ADRB1 is documented with different amino acids in the NCBI and UniProt databases.

The selected nsSNPs in each dataset were mapped onto protein complexes in the human 3D PPIN. The number of nsSNPs mapped on the 3D PPIN is given in Table 3.6*d*. Those nsSNPs were further classified by the regions where they were occurring, which are surface, interface, core and disordered regions. The number of nsSNPs in each region is listed in Table 3.1.

The hypothesis is that the larger the region, the higher number of mutations can be found at this site. In order to compare the propensity of nsSNPs amongst regions of occurrence, including nsSNPs$_{surface}$, nsSNPs$_{interface}$, nsSNPs$_{core}$ and nsSNPs$_{disordered}$, without bias due to their intrinsically different size, the number of classified nsSNPs by region was normalised by the corresponding size of the region. The following formula was used to analyse the enrichment of nsSNPs at each protein region.

$$P(SNP_{region}) = \frac{(SNP_{region}/size_{region})}{(SNP_{protein}/size_{protein})} \tag{3.2}$$

To avoid the possibility that the observed propensities were biased by a few proteins with large number of nsSNPs, confidence intervals (CI) at 95% were calculated by bootstrap re-sampling with 10,000 replicates. The statistical significance of differences between the propensities was estimated with pair-wise Wilcoxon comparison tests. These re-sampling and statistical analyses were performed with R [211]. These statistical analyses were also used for the nsSNP enrichment assessment under different criteria described in the following sections. Detailed statistical analyses are given in section "3.4.4 Statistical Evaluation".

The interface disease-related nsSNPs were further classified into inter-interface specific (nsSNPs$_{S_Inter}$), intra-interface specific (nsSNPs$_{S_Intra}$) and others (nsSNPs$_{Multi}$). An in-

	Inter-interface ($nsSNP_{S\_Inter}$)	Intra-interface ($nsSNP_{S\_Intra}$)
$nsSNP^{GD}$	256	500
$nsSNP^{SC}$	2,411	1,888

Table 3.7: Numbers of co-localised interface nsSNPs.

house written script was used to identify SNPs classes $nsSNP_{S\_Inter}$ and $nsSNP_{S\_Intra}$. A SNP is recognised as $nsSNP_{S\_Inter}$ by the script if at least one other interface nsSNP is located on the shared interface of the interaction protein pair considered and is involved in the same disease. On the other hand, a SNP is recognised as $nsSNPs_{S\_Intra}$ if at least one other interface nsSNP is co-localised at the same interface of a protein in the considered complex and is involved in the same type of disease. The number of nsSNPs in class $nsSNP_{S\_Inter}$ and $nsSNP_{S\_Intra}$ is listed in Table 3.7.

3.4.4 Statistical Evaluation

In this study, we are interested in characterising the features of disease-related mutations, particularly in identifying the tendency of mutations to occur at specific protein regions. The enrichment of mutations at each protein region was estimated by calculating the propensity of mutation-occurring residues at the protein region (surface, interface, core or disordered). A propensity which is larger than 1 (0 after logarithms) indicates that mutations occur frequently in a studied region. The propensities were also used to compare the relative enrichment of mutations between different protein regions. The statistical analyses were performed using R [211]. The details of the statistical analysis used in this study are described as follows.

Enrichment of Mutations

Assuming that mutations occur at random over protein sequences, we would expect to find a greater number of mutations at a larger protein region (the region containing the higher number of residues) than at a smaller protein region by chance. For instance, we expect to find more mutations at protein surfaces than at interfaces. Therefore, to compare the mutation enrichment between protein regions, the mutation numbers are normalised by the size (number of residues) of the region. This was calculated using the formula (Formula 3.2) given in the previous section "Data Collection and Analysis of SNPs" to obtain the propensities of the mutations in different protein regions.

Re-sampling

In order to estimate the observational bias induced by intensively studied proteins, the bootstrapping method was used to randomly re-sample from the pool of human proteins that have structural and mutation information. 10,000 replicates were generated for each pre-defined protein region with a mutation dataset (nsSNPs^{C}, nsSNPs^{GD} or nsSNPs^{SC}). 95 percentage confidence intervals were calculated from bootstrapping distributions. The density plots were generated to show the distributions of re-sampled datasets from the bootstrapping method. The **boot** function from the R programming language was used to perform the data re-sampling.

Statistical Significance

Statistical differences between propensities were estimated using the Wilcoxon Mann-Whitney test. This is a non-parametric test and assesses whether two populations have

an identical distribution. The **wilcox.test()** function from the R programming language was used to perform null hypothesis test.

Enrichment Compared with the Control

The enrichment of disease-related mutations in comparison with the control mutation dataset at each protein region was estimated using the fold-change, which is determined by:

$$Fold \quad Change \quad = \quad \bar{x} \quad / \quad \bar{y} \tag{3.3}$$

where (1) y_1, y_2, ..., y_n is the control; and (2) $\bar{x}$ and $\bar{y}$ are the means of the propensities obtained from the two compared mutation datasets at a specific protein region.

3.4.5 Definition of SNPs Disease Category

A previous study by Goh *et al.* [10] classified the diseases from OMIM genetic variation records into 21 disease types (Appendix A). This is currently the only available annotation for categorising OMIM variants. In this study, each nsSNP^{GD} was assigned with a disease type according to this classification.

For the cancer nsSNP dataset, the cancer category of nsSNPs^{SC} was defined by the cancer primary type. The cancer types with less than 10 nsSNPs in the COSMIC dataset were filtered out, this resulted in a final list with a total of 27 cancer types (Appendix B).

3.4.6 nsSNPs at Secondary Structure Elements

The composition of secondary structure of a protein was defined using DSSP [200]. By giving a PDB structure file of a protein, DSSP assigns the most likely class of secondary structure to each residue of the protein with a character indicating the secondary structure that the residue is part of. For proteins that do not have resolved structures, homologous structures (from the PSI-BLAST results described in the previous section) were used for this analysis.

Based on secondary structure profiles, DSSP secondary structure assignments were grouped into two main classes defined by us as stiff and flexible regions. Stiff regions include helix (H, G, I) and strand (B, E) structures in the secondary structure profile, while flexible regions include loops and turns (T, S, L). The three classes of nsSNPs, surface, interface and core nsSNPs, were assigned with propensities at protein stiff and flexible secondary structures (SS) using the formula:

$$P(SNP^{SS}_{region}) = \frac{(SNP^{SS}_{region}/size^{SS}_{region})}{(SNP_{protein}/size_{protein})} \tag{3.4}$$

The propensity of nsSNPs at these two SS groups (stiff/flexible) was calculated for each nsSNP class and their pairwise comparison was performed using the previously described statistical assessment.

3.4.7 Functional Site nsSNPs

The functional specificity of disease-related SNPs was examined with a structure-based measurement by screening the nsSNPs that are at or close to protein functional sites or post-translational modification (PTM) sites.

To investigate the enrichment level of close-to-functional-site nsSNPs, the functional site annotation was first downloaded from UniProt. It includes active sites, metal binding sites, binding sites, cleavage sites, inhibitory sites, and breakpoint sites. 2,100 proteins in the 3D PPIN were annotated with functional site information. An in-house script was written to screen protein structure coordinates and detect residues that are close to functional sites in 3D space. The distance between a functional site residue and any other residue was calculated by measuring the distance between C_α atoms from the two residues. A threshold value of 8 Å was used to identify residues close to the functional site. The nsSNPs, whose positions overlapped with these selected close-to-functional-site residues were defined as close-to-functional-site nsSNPs and were analysed using the formula 3.5 and the statistical assessment.

$$P(SNP_{region}^{closeToFunctionalSites}) = \frac{(SNP_{region}^{closeToFunctionalSites}/size_{region})}{(SNP_{protein}/size_{protein})} \tag{3.5}$$

The same analyses were performed for PTM sites. The PTM site annotation for human proteins was obtained from UniProt and PTMcode [204] databases. 4,370 proteins in the 3D PPIN were annotated with PTM site information. The same distance threshold and procedure were used to identify nsSNPs that are at or close to PTM sites in 3D space and calculate the enrichment of nsSNPs that are at or close to PTM sites.

3.4.8 Amino Acid Change of nsSNPs

In order to measure the frequency of different types of amino acid changes induced by nsSNPs, the amino acids were classified as non-polar (G, A, I, L, M, F, W, V), polar (S, T, N, Q, C, Y), positively charged (R, H, K), and negatively charged (D, E) residues according to their physico-chemical properties. Proline (P) was not assigned to any of these four physico-chemical classes since any residue change to or from proline is considered as a drastic change, so changes including proline were defined as an independent class. Figure 3.6 shows all the possible changes between residue classes. Apart from the aforementioned amino acid changes involving proline, drastic amino acid changes include substitutions between non-polar and positively charged residues, non-polar and negatively charged residues, and between positively changed and negatively charged residues. The rest of amino acid changes are defined as moderate changes. The changes between amino acids from the same class were defined as moderate but not included in our analysis.

The frequency of amino acid change types was calculated relative to drastic and moderate changes. For example, the frequency of changes between non-polar and polar residues was calculated as follow:

$$Number_{moderate\_change\_sub\_grp}/Number_{moderate\_change}$$

3.4.9 Analysis of Functional Similarity of Interaction Protein Pairs

To measure the functional similarity between an interaction protein pair, the method TAS [212] was implemented to measure the GO term specificity for the two proteins. Human protein GO annotations were downloaded from UniProt-GOA [213]. Proteins in the 3D PPIN were annotated with Biological Process (BP) functional terms. The GO database

termdb was also downloaded from the Gene Ontology web server (date 31 August 2013) to build a GO functional tree of the 3D PPIN. TAS combines the notions of topological distance and the lowest common ancestor distance methods (original paper Yu *et al.* [212]). The Lowest Common Ancestor Node (LCAN) is the lowest level of GO term that a pair of interaction proteins shares in common. All the parents nodes of this LCAN are the common ancestor nodes of this interaction protein pair. In other words, each node in the tree is not only the lowest common GO term of the protein interaction pairs considered, but also the common GO term of all the interaction pair of all its children nodes (Figure 3.21). Therefore, the fewer number of interaction protein pairs the LCAN of a considered protein pair contains, the higher specificity the LCAN is for the protein pair. On the other hand, the protein pairs were annotated as not being functionally similar if their LCAN is the GO root node (node 0 in Figure 3.21). A probability measure of the functional similarity given by a LCAN was calculated as following:

$P = n/N$

where n is the number of protein pairs in the common ancestor node, and N is the total number of protein pairs in the GO functional tree of the human 3D PPIN.

To study the consequences of nsSNPs on protein functions, the functional similarities between interaction protein pairs were used to determine the association between the types of diseases the nsSNPs lead to and the functions of the affected proteins. The PPIs were classified as the ones in which both proteins are related to the same type of disease ($\text{PPIs}_{disease}$), the ones whose partners are not involved in the same disease (PPIs_{diff}), and the ones that are not related to any disease (PPIs_{non}). A protein pair, in which only one of them was disease-related, was classified as PPIs_{non}. The functional similarity according to this scheme was calculated for proteins that are related to germ-line diseases and to

cancer and compared over the three classes of PPIs.

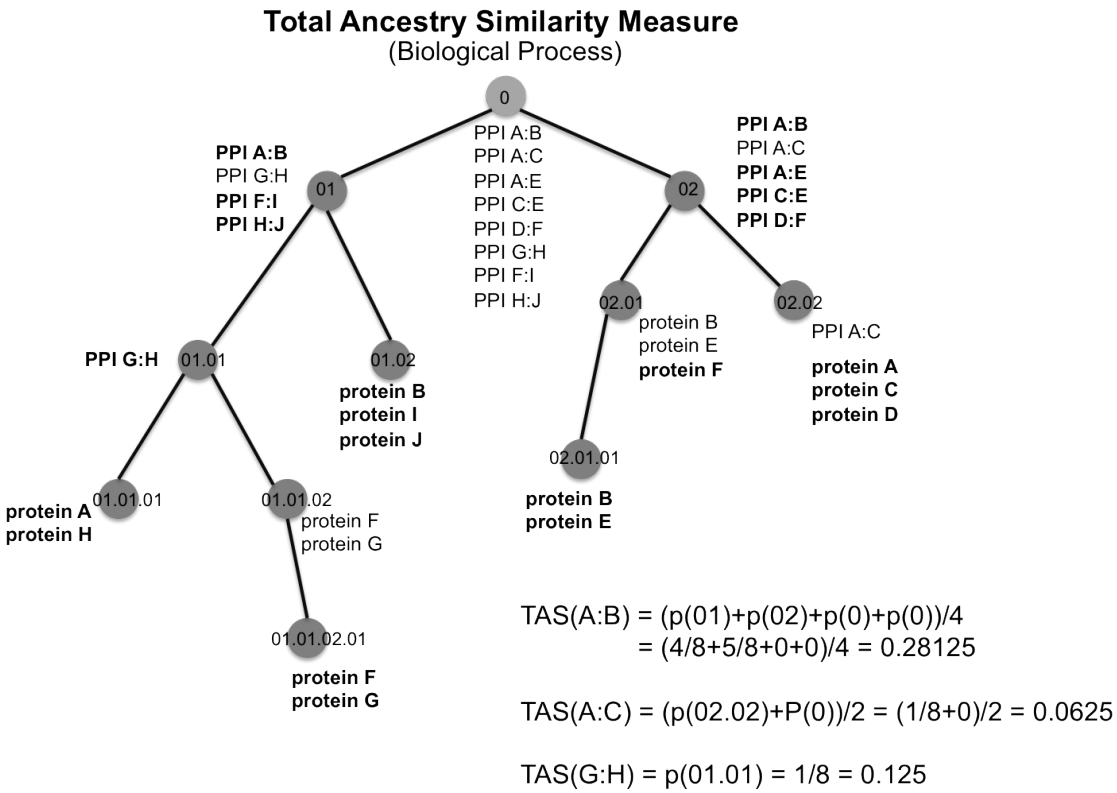

Figure 3.21: Total ancestry similarity measure. Nodes in the tree contain proteins which have the function of the node. Protein A, for example, has the function of the nodes "01.01.01" and "02.02". Nodes are also assigned with PPIs, when the node is the common functional ancestor of the protein pair. PPI A:B, for example, indicates the common functional ancestor nodes of Protein A and B. TAS(A:B) indicates the TAS score of protein A and B.

3.4.10 Function Annotation of SCOP Domain Superfamilies

Protein domains in the human 3D PPIN dataset were assigned with domain function categories using functional annotation of SCOP superfamilies [214, 215]. The domain function annotation (scop.annotation.1.73.txt and scop.larger.categories) was downloaded from the SCOP website. The nsSNPs^{GD} were mapped onto the function categories relative to the affected domains. The frequencies of nsSNPs^{GD} in each function category are shown in Figure 3.22.

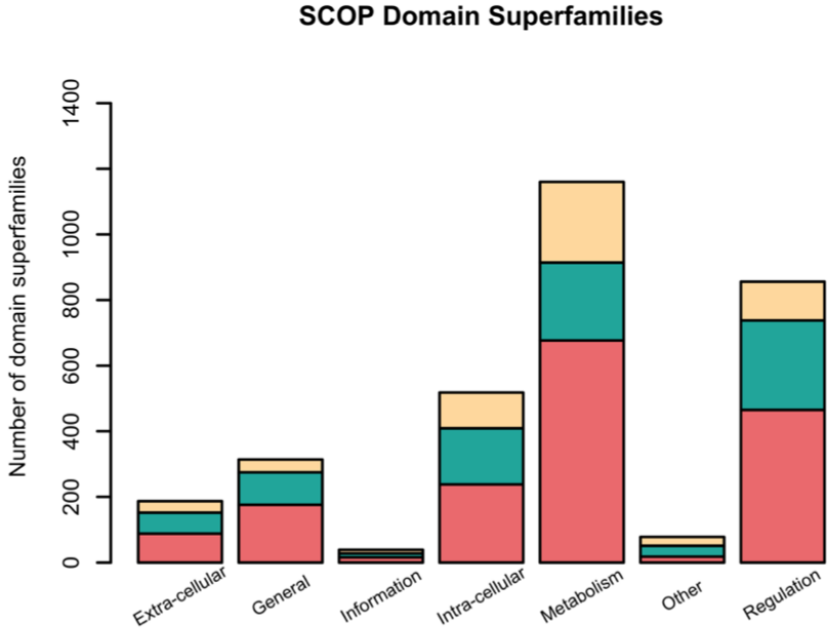

Figure 3.22: The functions of nsSNPs^{GD} occurring domain.

3.4.11 Prediction of nsSNP Impact

PolyPhen2 and Provean-1.1 [147] were installed on the local Unix machine to perform predictions in batch on nsSNP impact. The nsSNP^{GD} and nsSNP^{SC} datasets were analysed and compared between classes, including surface, interface, core and disordered.

3.4.12 Classification of Interface nsSNPs

Interface co-localised nsSNPs are classified into three classes: inter-interface specific (nsSNPs$_{S_Inter}$), intra-interface specific (nsSNPs$_{S_Intra}$) and others (nsSNPs$_{Multi}$). The implementation of interface nsSNP classification is presented as the following pseudo code.

Declare an array **SNP_arr** *store interface nsSNPs*
Set **n** *to be number of interface nsSNPs to be screened*
FOR *counter1 = 0 to n-1*:
    FOR *counter2 = 0 to n-1*:
        IF *SNP_arr[counter1] is not SNP_arr[counter2]*
          AND *SNP_arr[counter1] and SNP_arr[counter2] at interfaces of a protein pair*
          AND *SNP_arr[counter1] and SNP_arr[counter2] relate to the same disease*:
              IF *SNP_arr[counter2] on the opposing interface of SNP_arr[counter1]*
                AND *SNP_arr[counter1] not yet being reported as a nsSNP$_{Inter}$*:
                    report SNP_arr[counter1] as a nsSNP$_{Inter}$
              END IF

IF *SNP_arr[counter2] on the same interface of SNP_arr[counter1]*

AND *SNP_arr[counter1] not yet being reported as a* $nsSNP_{Intra}$**:**

report SNP_arr[counter1] as a $nsSNP_{Intra}$

END IF

ELSE**:**

report SNP_arr[counter1] as a $nsSNP_{Multi}$

END IF

END FOR

END FOR

Chapter 4

Pipeline to Generate 3D Protein-Protein Interaction Networks

Protein-Protein Interaction Networks (PPINs) have been largely used in the biological sciences to represent the gene/protein associations and the physical interactions occurring in the cell. By adding functional information to proteins within PPINs, the analysis can highlight hidden properties and assign more reliability to their association and the implied cellular processes. A detailed introduction of the PPIN applications and uses in biology has been presented in the first chapter. Recent studies have integrated structural information within PPIN in order to investigate the biological system at the molecular level and explore the underlying mechanisms at the atomic and molecular detail. In this chapter, the development of a pipeline for mapping 3D data to PPINs will be presented, as well as the comparison with the currently available 3D PPIN analyses.

4.1 Introduction

Over the last few years, analyses of 3D PPINs have gained momentum for their importance in studies such as understanding protein binding mechanisms, molecular properties of hub proteins, protein-protein interaction predictions, identifying druggable protein-protein interactions, and the impact of human genetic mutations on protein complexes. Several research groups have used 3D PPINs to conduct systematic studies on human disease-related mutations and revealed the molecular properties of those mutations [26, 27, 28, 91].

However, those studies have been limited by the relatively low availability of protein 3D structures. Homologous structures and structure modelling methods have been used to expand the 3D space of PPINs. Using a different approach, homologous proteins were used as structural proxies for the mapping of genetic variants and subsequent inference of complex interactions. In this chapter the automated pipeline for constructing 3D PPINs is presented, starting from the PPI data collection, searching for protein structures, the detection of interaction binding regions and the mapping of human mutation data onto these.

4.2 Materials and Methods

The automated pipeline of human 3D PPINs is constructed following a sequential process (Figure 4.1). Each protein in a 3D PPIN is annotated not only by its interaction relationship with other protein but also by the structural, functional and genetic information selected.

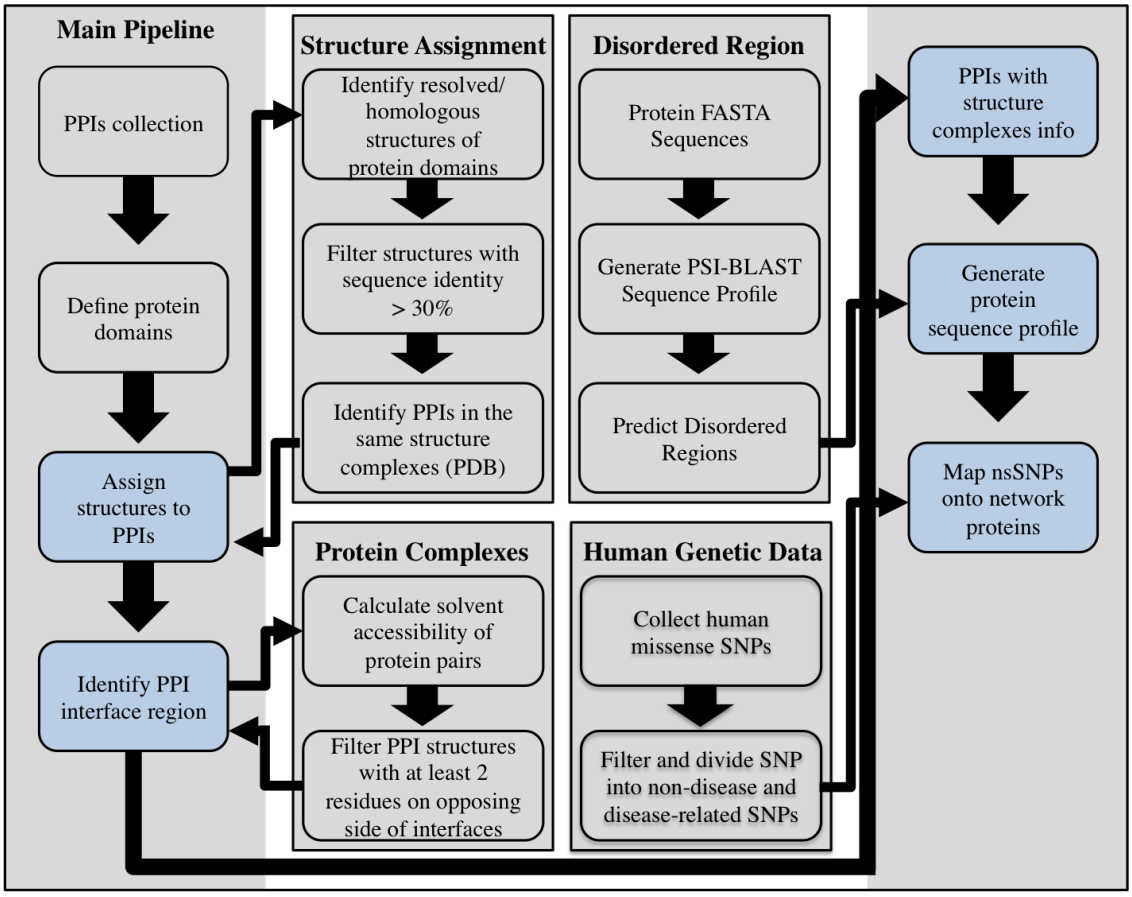

Figure 4.1: The automated pipeline for 3D protein-protein interaction network constructions. The processes coloured in light blue contain protein structure information.

Database (version)	Identifier	Num. of PPIs
BioGRID-ORGANISM-3.2.101	UniProt	209,838
DIP_20130131	UniProt/RefSeq	4,315
HPRD_Release9_062910	RefSeq, Gene symbol	39,240
IntAct (downloaded on 05 June 13)	UniProt	95,813
MINT_2012_02_06	entrez	33,954
STRING 9.05	Ensembl	4,445,596

Table 4.1: Datasets obtained from PPI databases. The number of human PPIs, the downloaded version/date, and the protein identifiers used in the databases.

4.2.1 Integration of Human Protein-Protein Interaction Datasets

Human PPI datasets were obtained from six publicly available databases, including BioGRID [60], DIP [61], HPRD [59], IntAct [62], MINT [63] and STRING [64]. The dataset from Havugimana *et al.* [67], which contains novel PPIs derived from AP-MS experiments, was also obtained. Only the protein interactions in which both are human proteins were selected. The initial number of PPIs from each database is reported in Table 4.1.

To integrate the PPI datasets, the pipeline only selects the protein interactions for which both proteins have an UniProt [203] reviewed accession identifier. The mapping between UniProt accession identifiers and other protein references are based on the identifier mapping provided by UniProt, which was downloaded from the UniProt FTP repository (ftp.uniprot.org). At this stage, the PPI dataset was not filtered with stringent criteria. However, the STRING database includes not only experimentally determined protein interactions, but also putative interactions obtained from functional association analysis such as genome context and co-expression analysis. In the downloaded STRING dataset (v9.05), only 6 % of PPIs/protein associations (263,666 out of 4,445,596) have experimental evidence of physical interactions. The putative PPIs can only indicate that both proteins of a PPI are a part of a biological pathway or have functions in common, but do not indicate direct physical interactions between associated protein pairs. The automated

	Total collected PPIs		Predicted PPIs	
Integrated from databases		1,878,098		1,689,416
Human 3D PPIN		39,387		25,773
	resolved	homologous	resolved	homologous
	4,423	34,964	657	25,117

Table 4.2: Number of PPIs obtained from STRING database. The numbers of initially collected PPIs and the PPIs in the 3D PPIN are given in the column "Total collected PPIs". These numbers includes the predicted PPIs obtained from STRING database. The column "Predicted PPIs" gives the numbers of PPIs obtained from STRING predicted PPI dataset and the PPIs in the 3D PPIN. The last row of the table gives the numbers of the PPIs which either have resolved structures or were assigned with homology models.

pipeline assigns structure complexes to PPIs that gives solid evidence of the physical interactions and reduce the false positive rate. The datasets which were integrated with the STRING dataset and without the STRING dataset both showed enrichment of disease-related mutations (nsSNPs^{GD} and nsSNPs^{SC}) at protein interface regions. A summary of the increases to the human PPIN by integrating STRING dataset is given in Table 4.2. By including the STRING predicted PPIs, the total PPIs increased to 1,878,098 interactions in the human PPIN.

4.2.2 Defining Protein Domains

The Protein Data Bank (PDB) [79] stores roughly around 5,000 human proteins in complexes. These are only about a quarter of estimated human proteins and many of them containing only partial protein structures. In order to construct the human 3D PPIN containing the largest number of protein structures, proteins are assigned with their domains boundaries.

A protein domain is a subunit of a protein structure which can fold independently and exist stably in isolation. Throughout evolution, functional domains shuffled and recombined to produce new proteins. Thus, the same domains can be found in different proteins, espe-

cially in eukaryotic organisms [83]. Finding homologous domain structures is an effective alternative for the 3D network construction.

Current available databases of protein domains include 3did [216], CATH [217], SCOP [218] and Pfam [219]. In this study, the Pfam domain sequence library was used to define protein domains for the following two reasons: Firstly, Pfam provides the profile Hidden Markov Models (HMMs) for the domain family, which are probabilistic models of seed alignments that enables fast and accurate sequence similarity searches using the HMM software HMMER3 [220]. Secondly, the output results from HMMER contain an e-value, which is the number of non-homologous hits in a specific database. The e-value allows users to filter the domain assignment with a given threshold.

The Pfam library and HMMER3 are embedded in the automated pipeline as follows:

1. Pfam 26.0 domain sequence library (Pfam-A.hmm), which was downloaded from Pfam FTP repository (http:// pfam.sanger.ac.uk/).

2. The sequence similarity-searching tool HMMER3 (http://hmmer.janelia.org/) was downloaded and installed on the local Linux machine.

3. Human protein sequences in FASTA format were downloaded from the UniProt FTP repository.

A

```
Query:       1-cysPrx_C  [M=40]
Accession:   PF10417.4
Description: C-terminal domain of 1-Cys peroxiredoxin
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence Description
    -------  ------ -----    -------  ------ -----   ----  --  -------- -----------
    4.3e-15   55.3   0.1    7.3e-15   54.6   0.0    1.4  1  Q06830
    2.7e-14   52.8   0.0    4.6e-14   52.0   0.0    1.4  1  P32119
    2.5e-13   49.7   0.1    5.1e-13   48.7   0.0    1.6  1  Q13162
      2e-12   46.8   0.1    3.4e-12   46.1   0.1    1.4  1  P30048
      5e-12   45.5   0.1    8.3e-12   44.8   0.1    1.4  1  P30041
  ------ inclusion threshold ------
       0.18   11.7   0.4       0.77    9.7   0.0    2.3  3  Q9UHY8


Domain annotation for each sequence (and alignments):
>> Q06830
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   54.6   0.0   2.4e-18   7.3e-15       1      40 []     162     197 ..     162     197 .. 0.97

  Alignments for each domain:
  == domain 1    score: 54.6 bits;  conditional E-value: 2.4e-18
                 HHH---TSS-B--TTT.T----BE---SSHHHHHHHHH-- CS
  1-cysPrx_C   1 ALQladkhgvatPAnWkPGdkvivpPppteeeavkryleg 40
                 A+Q++dkhg+++PA+WkPG ++i   +p+++  +k+y+++
      Q06830 162 AFQFTDKHGEVCPAGWKPGSDTI---KPDVQ-KSKEYFSK 197
                 9**********************...*****.9****985 PP

>> P32119
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   52.0   0.0   1.5e-17   4.6e-14       1      39 [.     161     195 ..     161     196 .. 0.95

  Alignments for each domain:
  == domain 1    score: 52.0 bits;  conditional E-value: 1.5e-17
                 HHH---TSS-B--TTT.T----BE---SSHHHHHHHHH- CS
  1-cysPrx_C   1 ALQladkhgvatPAnWkPGdkvivpPppteeeavkryle 39
                 A+Q++d+hg+++PA+WkPG ++i   +p ++  +k+y++
      P32119 161 AFQYTDEHGEVCPAGWKPGSDTI---KPNVD-DSKEYFS 195
                 9**********************...*****.7999997 PP
```

B

Domain_name	Pfam_id	Uniprot_id	Resi_start	Resi_end	E-value
CNH	PF00780	Q9Y4K4	516	816	7.5e-78
Pkinase	PF00069	Q9Y4K4	20	277	4.1e-65
PIP5K	PF01504	Q8TBX8	134	420	6.2e-74
C1_1	PF00130	P15056	235	280	2.7e-13
Pkinase_Tyr	PF07714	P15056	457	712	6.8e-67
RBD	PF02196	P15056	156	227	6.8e-32
LRR_4	PF12799	P39687	88	130	9.5e-11

Figure 4.2: Protein domain assignment. Examples of defined domains and the file data format. (A) Output result from HMMER. (B) The pipeline generated results extracted from the HMMER output file.

HMMER **hmmsearch** assigns domain definitions to query sequences of the human PPIN. The pipeline extracts the output from HMMER (Figure 4.2A) and generates a results file with format shown in Figure 4.2B. The results are then filtered with the following criteria: (a) The e-value of an assigned domain is smaller than 1e-3. (b) When two assigned domains of a protein overlap in the same region (sequence-based), the one with the lower e-value is selected. The total number of 16,471 proteins in the human PPIN were assigned with 4,706 different domains.

4.2.3 Sequence Alignment and Homologous Structure Detection

The identification of both experimental structure complexes and homologous complexes of PPIs is based on the output of the sequence alignment tool PSI-BLAST. The detailed implementation is described in this section.

The sequence alignment tools PSI-BLAST and T-Coffee were embedded in the automated pipeline as follows:

1. The local sequence alignment tool ncbi-blast-2.2.26+ [221] was downloaded from NCBI FTP repository (ftp.ncbi.nlm.nih.gov) and installed on the local Linux machine.

2. PDB biounit entries were downloaded from RCSB PDB FTP repository (ftp.wwpdb.org/pub/pdb/data/biounit/). The sequences of PDB structures were extracted and pre-formatted into a blast sequence library using the ncbi-blast-2.2.26+ **makeblastdb** command.

3. The profiles of PDB structure sequences were generated to record residue index information (Figure 4.3E).

4. The multiple sequence alignment tool T-Coffee [222] was downloaded and installed on the local Linux machine (T-COFFEE_distribution_Version_9.02.r1228.tar).

To identify resolved and homologous structures of proteins in the human PPIN, the BLAST command **psiblast** searches human protein sequences against the pre-compiled PDB sequence library with threshold e-value e-3 and 3 iterations. The pipeline extracts the information of hits and the sequence alignments from blast results (Figure 4.3A), and generates an output file **Human.blastAlignPdbStrucutre** containing all the blast results (one example shown in Figure 4.3B).

To optimise the sequence alignments between human proteins and PDB structure sequences, T-Coffee is implemented to re-align the sequence hits from PSI-BLAST (Figure 4.3C). Based on the domain assignments from HMMER, the pipeline extracts the alignments for the selected domain regions (Figure 4.3D). The following rules are applied to select the structures that cover the domain regions:

- The detected structure of a protein domain covers more than 80% of the protein domain sequence.
- The sequence identities of domain region alignments are calculated as the number of identical amino acids between two sequences divided by the length of the domain. The pipeline selects domain structures which have sequence identity higher than 30%.

The selected domain sequential and structural information are recorded in the file **Human.domStructure** and are used to construct 3D PPINs, which is described in the next section.

A

```
Query= 014514

Length=1584

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

lcl|4dlo_A  unnamed protein product                                    380    3e-118
lcl|4dlo_B  unnamed protein product                                    369    1e-114
lcl|1w0s_C  unnamed protein product                                    131    4e-31
lcl|1w0s_B  unnamed protein product                                    131    4e-31
lcl|1w0s_A  unnamed protein product                                    131    4e-31
lcl|1w0r_B  unnamed protein product                                    131    4e-31
lcl|1w0r_A  unnamed protein product                                    131    4e-31
lcl|3r6b_A  unnamed protein product                                    97.8   3e-22
lcl|1ls1_A  unnamed protein product                                    94.7   3e-21
lcl|4dlq_A  unnamed protein product                                    60.1   2e-08
lcl|3ghm_A  unnamed protein product                                    52.4   6e-06
lcl|3ghn_A  unnamed protein product                                    52.0   6e-06
lcl|4e0s_B  unnamed protein product                                    51.6   2e-05
lcl|4e5w_B  unnamed protein product                                    51.6   2e-05
lcl|3t5o_A  unnamed protein product                                    51.2   2e-05


>lcl|4dlo_A
Length=351

 Score =  380 bits (975),  Expect = 3e-118, Method: Compositional matrix adjust.
 Identities = 182/361 (50%), Positives = 256/361 (71%), Gaps = 13/361 (4%)

Query  575  PEPHEICDEDNFGAVIWKETPAGEVAAVRCPRNATGLILRRCELDEEGIAYWEPPTYIRC  634
            P P+EIC ED   +++WK TPAG++A  +CP NATG   RRC L   G+A+WE P++ RC
Sbjct  1    PAPYEICPEDYLMSMVWKRTPAGDLAFNQCPLNATGTTSRRCSLSLHGVAFWEQPSFARC  60

Query  635  VSIDYRNIQMMTREHLAKAQRGLPGEGVSEVIQTLVEISQDGTSYSGDLLSTIDVLRNMT  694
            +S +YR++Q   +EHLA   R L G+G+S+V +TL++++Q    Y+GDLL ++++LRN+T
Sbjct  61   ISNEYRHLQHSIKEHLA----RMLAGDGMSQVTKTLLDLTQRKNFYAGDLLMSVEILRNVT  117

Query  695  EIFRRAYYSPTPGDVQNFVQILSNLLAEENRDKWEEAQLAGPNAKELFRLVEDFVDVIGF  754
            + F+RA Y P    VQNF QI+SNLL EEN++KWE+AQ   P + EL +++EDF+ ++G
Sbjct  118  DTFKRASYIPASDGVQNFFQIVSNLLDEENKEKWEDAQQIYPGSIELMQVIEDFIHIVGM  177

Query  755  RMKDLRDAYQVTDNLVLSIHKLPASGA-TDISFPMKGWRATGDWAKVPEDRVTVSKSVFS  813
             M D +++Y +T N+V SI KLPA+   TDI+FPMKG +   DWA+  EDRV + KS+F
Sbjct  178  GMMDFQNSYLMTGNVVASIQKLPAASVLTDINFPMKGRKGMVDWARNSEDRVVIPKSIF-  236

Query  814  TGLTEADEASVFVVGTVLYRNLGSFLALQRNTTVLNSKVISVTVKPPPRSLRTPLEIEFA  873
            T ++  DE+SVFV+G VLY+NL    L   RN TV+N+SK+I VT++P P++  + LEIE A
Sbjct  237  TPVSSLDESSVFVLGAVLYKNLDLILPTLRNYTVINSKIIVVTIRPEPKTTDSFLEIELA  296

Query  874  HMYNGTTNQTCILWDETDVPSSSAPPQLGPWSWRGCRTVPLDALRTRCLCDRLSTFAILA  933
            H+ NGT N  C+LWD+++         LG WS +GC+TV  DA  T+CLCDRLSTFAILA
Sbjct  297  HLANGTLNPYCVLWDDSE---------SLGTWSTQGCKTVLTDASHTKCLCDRLSTFAILA  348

Query  934  Q  934
            Q
Sbjct  349  Q  349
```

B

```
>>      014514  575     934     4dlo_A  1       349     3e-118
ptn:PEPHEICDEDNFGAVIWKETPAGEVAAVRCPRNATGLILRRCELDEEGIAYWEPPTYIRCVSIDYRNIQMMTREHLAKAQRGLPGEGVSEVIQTLVEISQDGTSYSGDLLSTIDVLRN
MTEIFRRAYYSPTPGDVQNFVQILSNLLAEENRDKWEEAQLAGPNAKELFRLVEDFVDVIGFRMKDLRDAYQVTDNLVLSIHKLPASGA-TDISFPMKGWRATGDWAKVPEDRVTVSKSVFS
TGLTEADEASVFVVGTVLYRNLGSFLALQRNTTVLNSKVISVTVKPPPRSLRTPLEIEFAHMYNGTTNQTCILWDETDVPSSSAPPQLGPWSWRGCRTVPLDALRTRCLCDRLSTFAILAQ
pdb:PAPYEICPEDYLMSMVWKRTPAGDLAFNQCPLNATGTTSRRCSLSLHGVAFWEQPSFARCISNEYRHLQHSIKEHLA---RMLAGDGMSQVTKTLLDLTQRKNFYAGDLLMSVEILRN
VTDTFKRASYIPASDGVQNFFQIVSNLLDEENKEKWEDAQQIYPGSIELMQVIEDFIHIVGMGMMDFQNSYLMTGNVVASIQKLPAASVLTDINFPMKGRKGMVDWARNSEDRVVIPKSIF-
TPVSSLDESSVFVLGAVLYKNLDLILPTLRNYTVINSKIIVVTIRPEPKTTDSFLEIELAHLANGTLNPYCVLWDDSE---------SLGTWSTQGCKTVLTDASHTKCLCDRLSTFAILAQ
---
```

C

```
CLUSTAL W (1.83) multiple sequence alignment

O14514_575_934  PEPHEICDEDNFGAVIWKETPAGEVAAVRCPRNATGLILRRCELDEEGIAYWEPPTYIRCVSIDYRNIQMMTREHLAKAQ 80
4dlo_A_1_349    PAPYEICPEDYLMSMVWKRTPAGDLAFNQCPLNATGTTSRRCSLSLHGVAFWEQPSFARCISNEYRHLQHSIKEHLAR-- 78
                * *:*** ** : :::**.****::*  :** ****   ***.*. .*:*:** *:: **:* :**::*   :****:

O14514_575_934  RGLPGEGVSEVIQTLVEISQDGTSYSGDLLSTIDVLRNMTEIFRRAYYSPTPGDVQNFVQILSNLLAEENRDKWEEAQLA 160
4dlo_A_1_349    -MLAGDGMSQVTKTLLDLTQRKNFYAGDLLMSVEILRNVTDTFKRASYIPASDGVQNFFQIVSNLLDEENKEKWEDAQQI 157
                  *.*:*:*:* :**::::*  . *:**** ::::***:*: *:** * *:...****.**:**** ***:***:**

O14514_575_934  GPNAKELFRLVEDFVDVIGFRMKDLRDAYQVTDNLVLSIHKLPAS-GATDISFPMKGWRATGDWAKVPEDRVTVSKSVFS 239
4dlo_A_1_349    YPGSIELMQVIEDFIHIVGMGMMDFQNSYLMTGNVVASIQKLPAASVLTDINFPMKGRKGMVDWARNSEDRVVIPKSIFT 237
                  *.: **::::***:.::*: * *:::: :*.*:* **:****:    ***.***** :.  ***: .****.:.**:*:

O14514_575_934  TGLTEADEASVFVVGTVLYRNLGSFLALQRNTTVLNSKVISVTVKPPPRSLRTPLEIEFAHMYNGTTNQTCILWDETDVP 319
4dlo_A_1_349    -PVSSLDESSVFVLGAVLYKNLDLILPTLRNYTVINSKIIVVTIRPEPKTTDSFLEIELAHLANGTLNPYCVLWDDSE-- 314
                  :: **:****:*:***:**. :*.  ** **:***:* **::* *::  : ****:**: *** *  *:***:::

O14514_575_934  SSAPPQLGPWSWRGCRTVPLDALRTRCLCDRLSTFAILAQ 360
4dlo_A_1_349    ------SLGTWSTQGCKTVLTDASHTKCLCDRLSTFAILAQ 349
                      .**.** :**:**  ** :*:**************
```

D

```
HRM     PF02793 O14514  579     636     7.9e-09 4dlo_A  512     569     3e-118  51
>ptn:EICDEDNFGAVIWKETPAGEVAAVRCPRNATGLILRRCELDEEGIAYWEPPTYIRCVS
>pdb:EICPEDYLMSMVWKRTPAGDLAFNQCPLNATGTTSRRCSLSLHGVAFWEQPSFARCIS
DUF3497 PF12003 O14514  654     870     8.6e-45 4dlo_A  588     805     3e-118  48
>ptn:QRGLPGEGVSEVIQTLVEISQDGTSYSGDLLSTIDVLRNMTEIFRRAYYSPTPGDVQNFVQILSNLLAEENRDKWEEAQLAGPNAKELFRLVEDFVDVIGFRMKDLR
DAYQVTDNLVLSIHKLPAS-GATDISFPMKGWRATGDWAKVPEDRVTVSKSVFSTGLTEADEASVFVVGTVLYRNLGSFLALQRNTTVLNSKVISVTVKPPPRSLRTPLEI
>pdb:-RMLAGDGMSQVTKTLLDLTQRKNFYAGDLLMSVEILRNVTDTFKRASYIPASDGVQNFFQIVSNLLDEENKEKWEDAQQIYPGSIELMQVIEDFIHIVGMGMMDFQ
NSYLMTGNVVASIQKLPAASVLTDINFPMKGRKGMVDWARNSEDRVVIPKSIFT-PVSSLDESSVFVLGAVLYKNLDLILPTLRNYTVINSKIIVVTIRPEPKTTDSFLEI
GPS     PF01825 O14514  881     932     3.8e-12 4dlo_A  816     862     3e-118  53
>ptn:NQTCILWDETDVPSSSAPPQLGPWSWRGCRTVPLDALRTRCLCDRLSTFAIL
>pdb:NPYCVLWDDSE---------SLGTWSTQGCKTVLTDASHTKCLCDRLSTFAIL
```

E

```
>201l_A
M-1,N-2,I-3,F-4,E-5,M-6,L-7,R-8,I-9,D-10,E-11,G-12,L-13,R-14,L-15,K-16,I-17,Y-18,K-19,D-20,T-21,E-22,G-23,Y-24,Y-25,T-26,I
-27,G-28,I-29,G-30,H-31,L-32,L-33,T-34,K-35,S-36,P-37,S-38,L-39,N-40,A-41,A-42,K-43,S-44,E-45,L-46,D-47,P-48,A-49,I-50,G-5
1,R-52,N-53,T-54,N-55,G-56,V-57,I-58,T-59,K-60,D-61,E-62,A-63,E-64,K-65,L-66,F-67,N-68,Q-69,D-70,V-71,D-72,A-73,A-74,V-75,
R-76,G-77,I-78,L-79,R-80,N-81,A-82,K-83,L-84,K-85,P-86,V-87,Y-88,D-89,S-90,L-91,D-92,A-93,V-94,R-95,R-96,A-97,A-98,L-99,I-
100,N-101,M-102,V-103,F-104,Q-105,M-106,G-107,E-108,T-109,G-110,V-111,A-112,G-113,F-114,T-115,N-116,S-117,L-118,R-119,M-12
0,L-121,Q-122,Q-123,K-124,R-125,W-126,D-127,E-128,A-129,A-130,V-131,N-132,L-133,A-134,K-135,S-136,R-137,W-138,Y-139,N-140,
Q-141,T-142,P-143,N-144,R-145,A-146,K-147,R-148,V-149,I-150,T-151,T-152,F-153,R-154,T-155,G-156,T-157,W-158,D-159,A-160,Y-
161,K-162,
```

Figure 4.3: Protein local sequence search. Example of PDB structure searching results and the file format. (A) Output result from NCBI PSI-BLAST. (B) The pipeline generated results extracted from PSI-BLAST output file. (C) The sequence alignment between a protein sequence and PDB sequence from T-Coffee. The coloured regions indicate domains listed in (D). (E) PDB structure sequence profile.

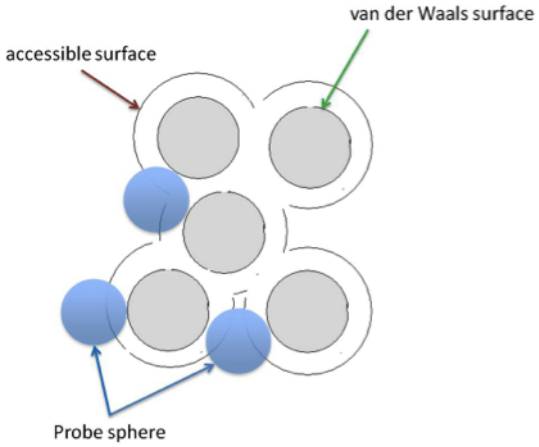

Figure 4.4: Molecular solvent accessible surface area.

4.2.4 3D Interaction Network Construction

Having identified protein domain structures, the pipeline constructs the human 3D PPIN. Firstly, the pipeline selects PPIs whose protein pairs are contained in the same PDB entry. When more than two PDB entries are found for a interaction protein pair, the PDB entry with the highest sequence identity to the protein pair, is selected. The total of 9,091 proteins in the human PPIN were found to have either resolved or homologous structures in the same PDB entries with their interaction protein partners.

To determine the physical interactions of protein pairs, which have domain structures in the same PDB entry, the tool POPSCOMP [102] was used to generate the profiles of normalised Solvent Accessible Surface Area (SASA) (Figure 4.4) and the buried area upon complex formation of protein pairs. A surface residue is considered to be part of a protein interface region if it buries more than 15% of its SASA upon complex formation. To ensure the physical interaction of proteins pairs, the pipeline selects the PPIs if at least 2 interface residues from each protein of a interaction pair are identified.

The interaction protein pairs for which homologous structures were used for the identifications of binding regions were mapped with the relative position using sequence alignments.

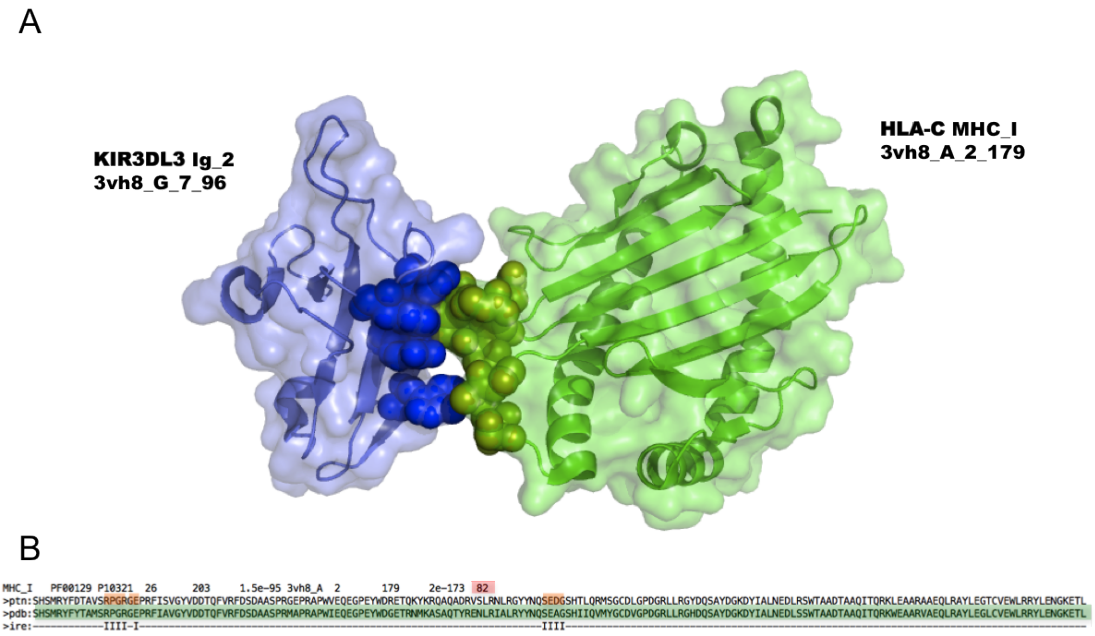

Figure 4.5: Mapping the relative position of protein interface region. (A) The homologous structures of protein HLA-C and KIR3DL3. The regions shown in solid colour with sphere representation indicate the binding regions of the two proteins. (B) The sequence alignment between the homologous (PDB 3vh8 chain A) and protein HLA-C. The pipeline generates the file recording sequence alignments of proteins and their homologous structures, and related information. The first line states the domain information of the protein and the value of the sequence identity (highlighted in red). **>ptn:** indicates the protein sequence. **>pdb:** indicates the structure sequence. **>ire:** indicates the residues that were identified to be at interface region of the structure. The putative interface residues are highlighted in colour orange.

An example shown in Figure 4.5, an homologous structure (PDB entry 3vh8) of protein HLA-C domain MHC_I was used to calculate the SASA values of residues and to identify the residues that have the physical contact with the partner protein KIR3DL3. In Figure 4.5, the region shown in solid green sphere representation indicates the binding region of the homologous structure of HLA-C. The relative binding region of HLA-C was identified through the sequence alignment.

The resulting 3D PPIN contained 8,249 proteins with 39,387 interactions. Each protein was annotated with structure regions, including surface, interface and core using the method described in this section.

4.2.5 Inter-domain Disordered Region Prediction

The importance of protein disordered regions in mediating protein interactions and functions were described in chapter 2. In this project, the disordered region is defined as structural elements outside of domains and predicted to have a functional role by the predictor DISOPRED [176]. These flexible structural regions are distinctive to loop secondary structures. The predictor DISOPRED is embedded into the pipeline to predict the disordered regions of proteins in the 3D PPIN. The implementation is described in the following:

1. blast-2.2.26 was downloaded and installed on the local Linux machine.
2. BLAST non-redundant sequence database was downloaded (ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.*.tar.gz).
3. The sequence profiles of the proteins in the human 3D PPIN are generated using PSI-BLAST to search against the non-redundant sequence database.

The protein sequence profiles of proteins generated from PSI-BLAST are used by the pipeline to generate a symmetric vector of residues. DISOPRED [176] takes the residue vector as input file for protein disordered regions prediction.

4.2.6 Protein Sequence Profile

The pipeline generates the sequence profile of proteins in the human 3D PPIN. The profile documents different protein regions defined in the last two sections, including surface, interface, core and disordered regions. One example shown in Figure 4.6.

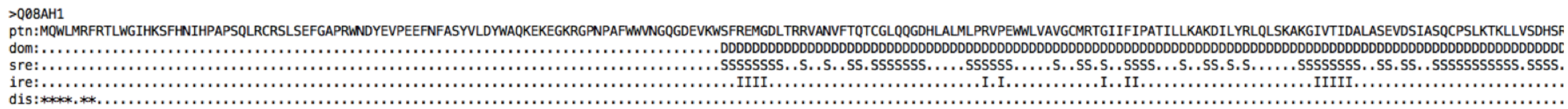

Figure 4.6: Protein sequence profile. An example in the protein sequence profile. The profile of a protein starts with >(protein UniProt accession identifier) **ptn:** indicates the protein sequence. **dom:** indicates the domain regions with **D** character. The domain region assignment were obtained from sequence similarity searching of the Pfam-A library. **sre:** indicates the solvent-exposed residues with **S** character. **ire:** indicates the binding region residues with **I** character. **dis:** indicates disordered region residues with * symbol.

4.2.7 Mapping of SNP Data

As described in chapter 3, three human genetic variant datasets were obtained from databases including, dbSNP, OMIM and COSMIC. They are representing different types of genetic variation: general variants, germ-line disease-related variants, and somatic cancer variants respectively. The occurring positions of SNPs are also annotated in the sequence profile as shown in Figure 4.7.

Each group of nsSNPs are classified by the occurrences of the protein regions. The sequence profile is used to divide the nsSNPs. A nsSNP is classified as a surface nsSNP if the relative position of **sre:** sequence is annotated as **S** and the relative position of **ire:** sequence is annotated as **-**. Whereas, a nsSNP is classified as an interface nsSNP if the relative position of **ire:** sequence is annotated as **I**. The remaining residues at domain regions with annotations of **D** symbol on **dom:** sequence, and annotations of **-** symbol on both **sre:** and **ire:** sequences are defined as core region residues. The nsSNPs occurring at those residues are classified as core nsSNPs.

Moreover, the residues annotated with a ***** symbol on **dis:** sequence and **-** symbol on **dom:** sequence are defined as disordered region residues. The occurrences of nsSNPs at those residues are classified as disordered region nsSNPs.

The total number of nsSNPs in each class and a detailed analysis were presented in the previous chapter.

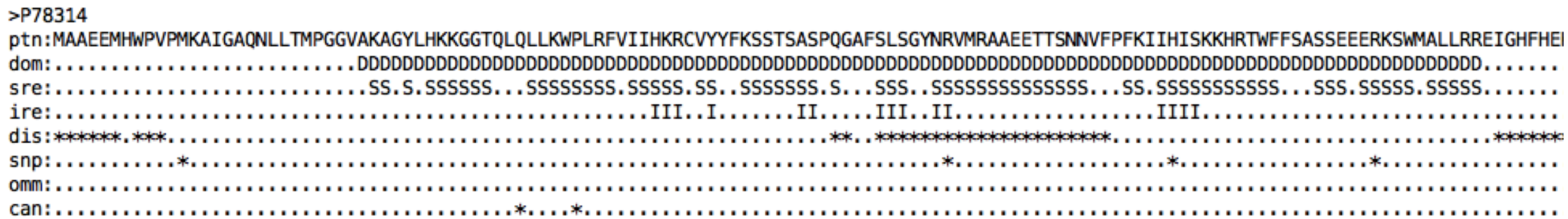

Figure 4.7: Protein sequence profile with SNP annotation. An example in the protein sequence profile. **snp:** indicates the positions of common nsSNPs with * symbol. **omm:** indicates the positions of OMIM nsSNPs with * symbol. **can:** indicates the positions of cancer (COSMIC) nsSNPs with * symbol.

4.2.8 Comparison with Other Existing Applications

Two recent studies [27, 24] presented human 3D network development and the implication of the networks. In this section, a summary of the comparison with those two studies is presented in Table 4.3.

Our approach is substantially different in:

1. Being able to start from PPIN and automatically extract the Common Variants, and disease-related nsSNPs onto different regions: Surface, Core, Interface and inter-domain Disordered region. To our knowledge, no existing method is offering this.

2. We extract the SNP occurrences and propensities by mapping onto human 3D structures and corresponding positions on relative homologous structures. This allows us to expand the existing sequence-3D gap and enrich our data (total number of 3D PPIN)

3. We extract nsSNPs co-localised on the same and opposing interfaces.

	Wang *et al*	**Interactome3D**	**this study**
Initial PPI data source (databases)	HPRD, BioGRID, IntAct, MINT, VisANT and iRefWeb	IMEx, BIND, BioGRID and HPRD	BioGRID, DIP, HPRD, IntAct, MINT, STRING and Havugimana *et al*
Protein structures	resolved structures of the two proteins in complex, or homologous mapping through iPfam	1) entirely or partially resolved protein structures; 2) homologous models from Modbase; 3) homologous structures; 4) domain-domain structural templates in 3did database. (The homologous structures and templates were used as templates to model the structures)	resolved structures or homologous structures with sequence identity >30%
Identification of interface regions	the corresponding interfaces given in 3did or iPfam databases	1) covalent interactions, defined as two sulfur atoms of a pair of cysteines at a distance $\leq$ 2.56 Å; 2) hydrogen bonds, defined as all atom pairs N-O and O-N at a distance $\leq$ 3.5 Å; 3) salt bridges, defined as all atom pairs N-O and O-N at a distance $\leq$ 5.5 Å; 4) van der Waals interactions, defined as all pairs of carbon atoms at a distance $\leq$ 5.0 Å.	POPSCOMP method was used to calculate the SASA and identify the residues that are at the binding regions
Numbers of proteins in the 3D PPIN	2,816	4,239	8,249
Human mutation data source	HGMD database and dbSNP (build 132)	N.A.	dbSNP (build 135), OMIM and COSMIC
Analysis of human mutations	1) enrichment of mutations and SNPs on interaction interfaces (comparing the observed number of mutations and SNPs on interfaces to the relative length of the protein sequence forming the interface); 2) pair-wise interface mutation calculations.	N.A.	1) enrichment of nsSNPs (missense variants) at predefined protein regions; 2) functional roles of nsSNPs; 3) pair-wise interface nsSNPs calculations.

Table 4.3: Comparison between 3D PPINs.

4.2.9 Applications

The applications of the automated pipeline will be presented in this section. Human PPI sub-networks were generated for the studies focused on different proteins, including B-RAF, hub proteins and LAMIN.

The B-RAF Mutants

The application summarised here is published in the paper of Satoh *et al.* [33], where more details about the studied system are given. Here I report my contribution to the work.

The mutations p.V600E, p.600DLAT and p.T599A (Figure 4.8) were found in Langerhans Cell Histiocytosis (LCH) patients. The automated pipeline developed in this thesis was used to search for domain definition and identify the homologous structures of human B-RAF. The protein was assigned with the domain RBD and Pkinase_Tyr by searching the Pfam-A domain library. The Pkinase_Tyr domain was further investigated as it is the domain in which those three mutants of our interests occur and was found with homologous structures. The PDB structure (1WUH) with the highest sequence identity to the human B-RAF protein sequence was used as the template to build the structure models of B-RAF mutants using the tool Modeller 9v8 [223].

Our results showed that the ^{600DLAT}B-RAF insertion have the same structural and functional effects as the ^{V600E}B-RAF mutant on B-RAF that destabilise the inactive conformation of the B-RAF kinase and consequently increase ERK activation. Whereas, the ^{T599A}B-RAF mutant was found to be a germ-line polymorphism and not having effect on the inactive conformation of the B-RAF kinase. Instead, T599A occurring at a major phosphorylation site of the B-RAF activation domain suppress B-RAF activity.

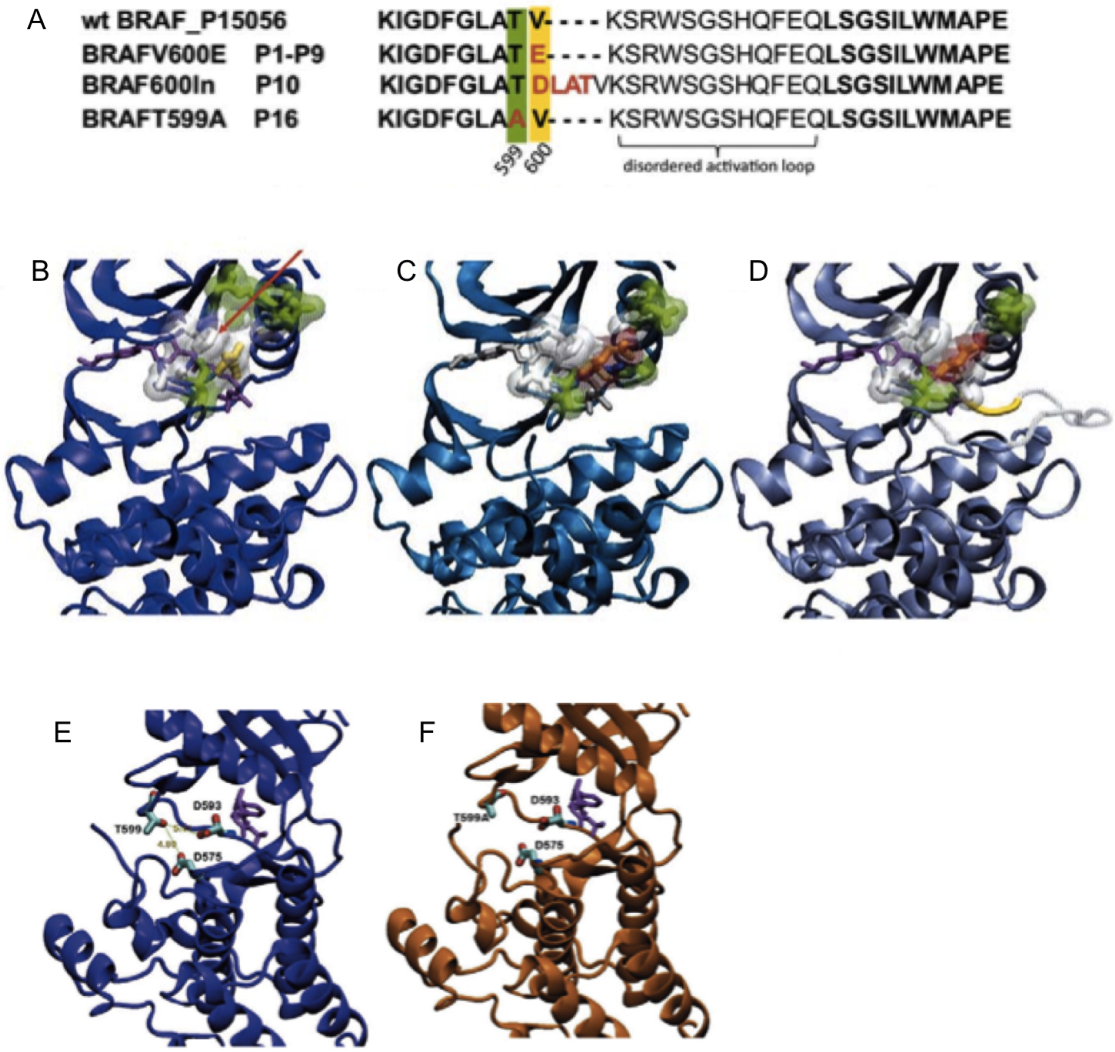

Figure 4.8: B-RAF kinase domain mutants. (A) Protein sequence alignments between reference (wild-type) sequence and patients' sequences. The second column indicates the patients' ID number. (B-D) Comparison between ^{wt}B-RAF ((B), purple), ^{V600E}B-RAF structure ((C), cyan) and the modelled mutant ^{600DLAT}B-RAF ((D), grey). (B) Val600 (yellow) forms a hydrophobic contact with Phe468 (red arrow). In (C) and (D) charged residues Asp and Glu (in orange) disrupt the hydrophobic network of interactions, stabilising the active conformation of the P-loop. (D) Insertion Asp-Leu-Ala-Thr shifted Val600 and disrupt the hydrophobic cluster. (E, F) Comparison between models of ^{WT}B-RAF ((E), violet) and ^{T599A}B-RAF ((F), gold). ^{T599A}B-RAF substitutes a polar uncharged residue with a hydrophobic residue, causing the loss of short-ranged interactions with residues D576 and D594. Figure adapted from Satoh *et al.* (2012) [33].

Promiscuous Residues of Hub Proteins

The application summarised here is published in the paper Fornili *et al.* [113], where more details about the studied system are given. Here I report my contribution to the work.

Promiscuous residues at the binding interfaces of hub proteins play essential role in the protein binding events. The flexibility of those residues enable the proteins to adopt different conformation according to the partner proteins.

In the study of Fornili *et al.*, two protein datasets were generated to study the properties of promiscuous residues: S^{Full} and S^{Soc}. S^{Full} is a non-redundant list of proteins generated from PiSite database, while S^{Soc} is composed of sociable proteins, defined as having at least 3 structural partners and 3 different binding states. The interface residues of proteins were defined with Binding Multiplicity (BM), which is the number of partner proteins that a residue involves in the binding interfaces. The interface residues were classified into three binding classes by BM value. Residues with BM $\geqslant 2$ were classified as c_{multi}, while residues with BM $= 1$ were classified as c_{mono} if belonging to monopartner proteins and as $c_{mono\_in\_multi}$ if belonging to multi-partner proteins.

The pipeline was used to map human nsSNPs onto the relative positions on the studied proteins in order to see whether promiscuous residues are prone to the mutations. Human homologous proteins of S^{Full} and S^{Soc} proteins were identified using NCBI-BLAST. Human nsSNP datasets were also obtained from dbSNP and OMIM, and defined as $nsSNPs^{C}$ and $nsSNPs^{GD}$, respectively. The position mapping between the studied proteins and human homologous relied on the sequence alignments from the BLAST results. A total of 38 S^{Full} proteins and 25 S^{Soc} proteins were annotated with nsSNP information.

Promiscuous positions in S^{Full} proteins (orange) were found less rich with $nsSNPs^{C}$ than

both classes of monopartner residues (cyan and blue). This observation was also found in S^{Soc} proteins with reduced statistic significance. This may suggest that the human equivalent of the promiscuous positions considered here tend to be less tolerant to genetic variation. The promiscuous residues are under higher level constraints in order to preserve effective binding. The occurrences of mutations at promiscuous positions may be prone to result in a lethal phenotype. On the other hand, the analysis of $nsSNPs^{GD}$ was strongly affected by the small number of observations, which requires a larger human protein dataset to obtain a more accurate investigation.

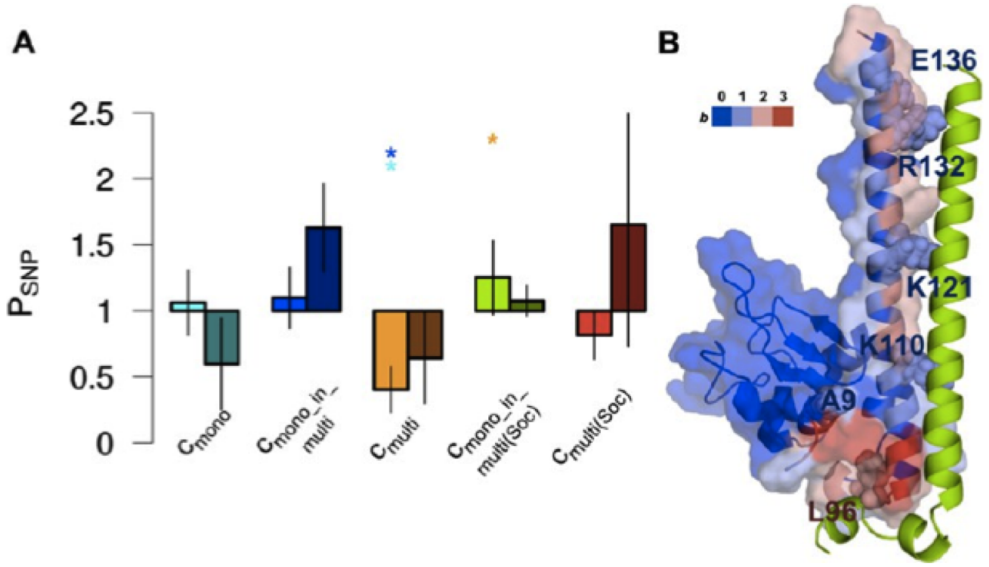

Figure 4.9: The occurrences of nsSNPs at promiscuous residues. Propensities of $nsSNPs^{C}$ and $nsSNPs^{GD}$ in S^{Full} and S^{Soc}. (A) Propensities of $nsSNPs^{C}$ (light colours) and $nsSNPs^{GD}$ (dark colours) relative to the interface residues of c_{mono} (cyan), $c_{mono_i n_m ulti}$ (blue), c_{multi} (orange), $c_{mono_i n_m ulti(Soc)}$ (green) and $c_{multi(Soc)}$ (red). The propensity is calculated per protein. The reported values are averages over S^{Full} monopartner proteins for c_{mono}, S^{Full} multipartner proteins for $c_{mono_i n_m ulti}$ and c_{multi}, and S^{Soc} proteins for $c_{mono_i n_m ulti(Soc)}$ and $c_{multi(Soc)}$. (B) SNPs in the human survivin protein. SNPs found in the interface region of survivin are labelled and represented as van der Waals spheres. A survivin binding partner (borealin) is also represented as green cartoon (PDB ID: 2RAW). Figure adapted from Fornili *et al.* (2013) [113].

The Immunoglobin-like Domain of LAMIN

Mapping pathogenic mutations in Laminopathies onto the Lamin Ig-like fold. Laminopathies are caused by mutations in A-type nuclear lamins. Although being quite accurately described and mapped onto the LMNA gene, it is difficult to extract precise structural information on the genotype-phenotype relation, due to the scarcity of resolved structures for this gene.

Only the Ig-like fold domain has been structurally resolved, and therefore we used available bioinformatics tools like PolyPhen-2, Fold X, Parameter OPtimised Surfaces, PocketPicker and the pipeline developed in this thesis to characterise 56 missense mutations for position, surface exposure, change in charge and effect on Ig-like fold stability. The striking result was that the majority of mutations (21/27) associated with a skeletal muscle phenotype mapped onto the Ig-like domain, are non-surface exposed and we predict them to affect the stability of the Ig-like fold domain. Interestingly, the other 6 mutations clustered together, showed increased surface exposure, and no effect on the protein stability.

We observed a clear separation (Figure 4.10A) of laminophaties specific phenotypic groups Skeletal muscle, Lipodystrophy and Premature ageing and clustering on different regions on the Surface/Core of the Ig-like fold. The relative positions matched different electrostatic surfaces, indicating a different possible impact of the mutations on the implicated molecular mechanisms and nature of the interaction partners. The Skeletal muscle mutations clustering in distinct, charged regions can affect lamin A/C -protein/DNA/RNA interactions may suggest a distinct pathological mechanism for this phenotype.

The pipeline was used to extract nsSNP information of lamin. One missense SNP (p.T534S) which is not known to be related to disease was found in the nsSNP dataset obtained from

dbSNP and was used as a control. This nsSNP was found to be in a region of the Ig-like fold domain but does not overlap with any of the clusters that contain pathogenic mutations (Figure 4.10B). Moreover, it was predicted to have a low $\Delta\Delta G$ by Fold X, and to be benign by PolyPhen2.

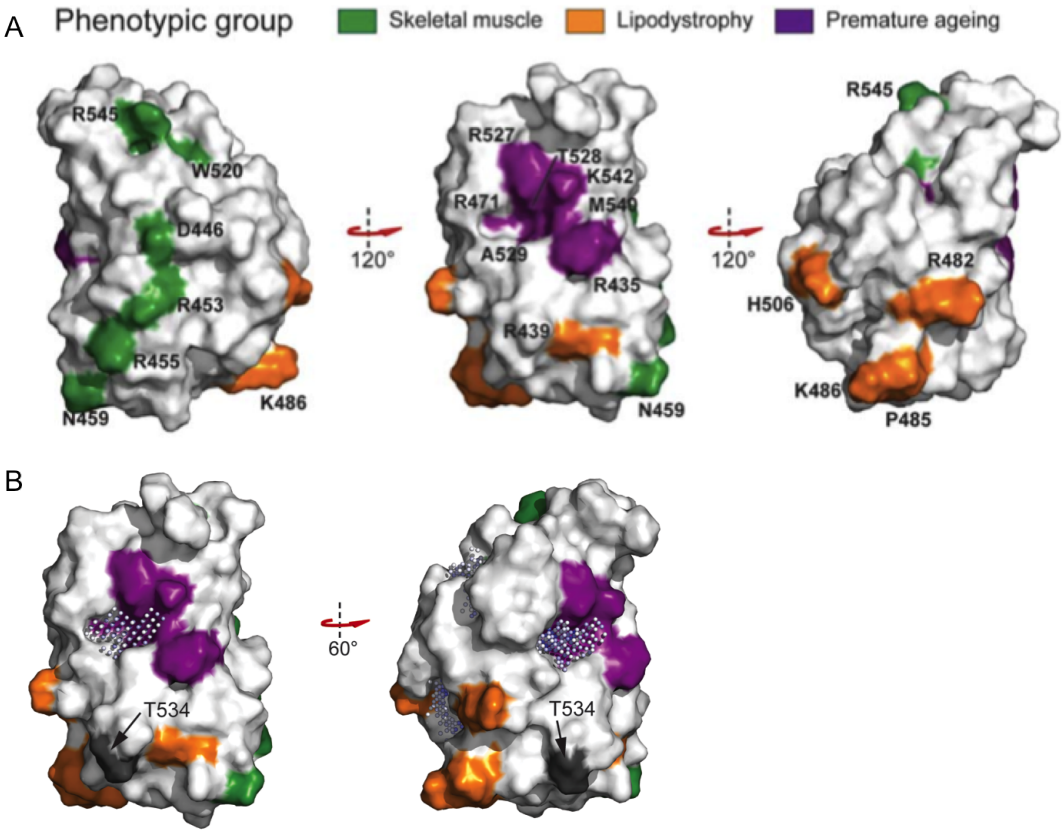

Figure 4.10: The occurrences of nsSNPs in LAMIN Ig-like fold domain. (A) The six skeletal muscle cluster residues (green) and all residues associated with premature ageing (purple) and lipodystrophy (orange) are highlighted in the surface representation, and can all be seen to form separate clusters. (B) Figure adapted from Scharner *et al.* (2013) [224].

Chapter 5

Summary and Future Direction

The usage of Protein-Protein Interaction Networks (PPINs) have progressed from merely abstract to detailed representation of interacting protein pairs annotated with functional and structural information. The structural information extracted from 3D PPINs enables the study, for example, of biological mechanisms occurring in the cell at the molecular detail, and can result very useful for the detection of druggable candidate PPIs for the clinical treatment. Efforts have been made to identify the features of human disease-related genetic variation using human 3D PPINs [26, 91, 27, 28]. It is commonly agreed that protein interfaces play an important role for protein binding activity. An interesting finding confirmed in this work is that mutations occurring at interface regions of protein complexes are more likely to result in disease than mutations occurring at other protein regions. This has confirmed previously observed enrichment of disease-related variants at interface regions [27].

However, the number of experimentally resolved human protein structures is comparatively small considering the number of known human proteins. Homologous structures

and structure modelling methods have been used to compensate the low number of experimentally determined structures and increase the 3D space of PPINs. In the study of Mosca *et al.* [24], homologous structures were selected with stringent criteria to build protein complex models. The structure complexes of domain-domain interactions were used to increase the size of human 3D PPIN [216]. This approach has significantly increased the number of structure complexes mapped to PPIs, with more than 4,000 proteins in the network compared to earlier work by Wang *et al.* [27] containing less than 3,000 proteins in the 3D network.

In this study, we aimed to enlarge the 3D space of PPINs to annotate most proteins in the network with functional, structural and genetic information. An automated pipeline was developed for generating human 3D PPINs. This has the practical advantages, that allows for a standardised procedure for 3D PPIN construction. It has been used here for human 3D network constructions and can be used to construct 3D PPINs of other organisms in the future. The flexibility of the pipeline also allows the implementation of studies on specific diseases with data generated from next generation sequencing or GWAS studies. Moreover, such a pipeline allows the user to assign the thresholds for homologous structure selection and interface residue selection. With a sequence identity target of more than 30% for homologous structure selection, the human 3D PPIN constructed for this study contains more than 8,000 proteins. Finally, the pipeline has the function of mapping available datasets of human genetic variation onto proteins constituting the 3D PPINs. The human 3D PPIN generated for this study was further mapped with non-synonymous Single Nucleotide Polymorphisms (nsSNPs) data obtained from databases including dbSNP, OMIM and COSMIC. The nsSNP dataset from dbSNP was used as the control dataset to be compared with disease nsSNPs from OMIM and COSMIC in

order to differentiate the properties between disease and non-disease related nsSNPs. This control dataset is noted as nsSNPs^{C}, and was filtered by omitting the nsSNPs from OMIM and COSMIC databases. The nsSNP datasets from OMIM and COSMIC represented the Germ-line Disease nsSNPs (nsSNPs^{GD}) and Somatic Cancer nsSNPs (nsSNPs^{SC}), respectively, and were analysed as two data groups.

The nsSNPs were classified and analysed by their occurrence in the defined protein regions, including surface, interface, core and inter-domain disordered regions. Our analyses showed an enrichment of nsSNPs^{GD} at the protein interface regions as the aforementioned study reported [27]. nsSNPs^{SC} were also found to be enriched at the interfaces (see Chapter 3 "Enrichment Analysis of Disease-related nsSNPs"). On the other hand, nsSNPs^{C} were found to be proportionally in high number at disordered regions, where both nsSNPs^{GD} and nsSNPs^{SC} were found under-represented. This observation may suggest that residues at disordered regions are under a lower number of structural constraints to preserve protein function and therefore can be more easily mutated and not result in diseased states.

In addition, the structural properties of nsSNPs were also analysed by looking at the preferences of nsSNPs in physicochemical properties and secondary structure elements. Our results showed that nsSNPs^{GD} exhibit distinctive structural features compared with the control nsSNPs^{C}. More than two-fold enrichment of nsSNPs^{GD} at interface regions was reported when compared with the propensity of nsSNPs^{C} at the same region. Moreover, nsSNPs^{GD} are prone to cause drastic changes in amino acid type which are more likely to affect protein stability and interactions with other proteins (in Chapter 3 "Distinguishable Structural Features of Disease-related nsSNPs").

nsSNPs^{SC} were found to share similarities with nsSNPs^{C} in terms of structure preferences and trends in physicochemical changes. However, the observation of nsSNPs^{SC} are most

likely dominated by passenger mutations which do not contribute to oncogenesis and are the large majority in the nsSNPs^{SC} dataset. It remains a challenge in cancer research to pin point driver mutations by sequencing a cancer genome.

The functional properties of nsSNPs were also investigated by looking at the propensities of nsSNPs which are spatially close to functional residues, including active sites, binding sites, PTM sites and other functional residue annotations from UniProt (Chapter 3 "Functional Specificity of nsSNPs"). nsSNPs were in general found not favoured to be close to the functionally important residues over all three SNP data groups: propensities (nsSNPs^{C}, nsSNPs^{GD} and nsSNPs^{SC}) are lower than 1. By comparing between non-disease and disease-related nsSNPs, we observe that both nsSNPs^{GD} and nsSNPs^{SC} have a higher propensity to be close to functional residues. This may suggest that the occurrences of variants close to functional residues are more likely to be lethal and therefore not observed in the population. If close-to-functional-residue mutations are transmitted through generations, they are more likely to have a critical impact on the protein function and subsequently lead to disease.

Another important contribution of the 3D PPIN studies is in highlighting PPI druggable targets [128, 225, 226]. In our study, a high number of nsSNPs^{GD} pairs occurring at the same interfaces were found to have the tendency to cause the same type of disease (Chapter 3 "Co-localised Disease-related nsSNPs"). The molecular information of the binding interface could help in designing compounds that are more specifically targeting a protein interface containing mutations relative to a specific disease, and have more effective therapies. Additionally, our results showed that interacting protein pairs which have the same biological function (GO annotation) have the tendency to cause to same types of germ-line disease. Our results support the idea that PPIs can potentially be ideal

drug targets for specific diseases caused by specific PPIs, even though it has been very challenging to find effective targets so far (the detailed discussion can be found in Cochran (2000) [225] and Higueruelo *et al.* (2013) [128]).

3D PPINs have shown to be genuinely an effective tool for implementing large-scale studies on biologically relevant data. Our results showed distinctive structural features of germ-line disease SNPs from other genetic variants. A comprehensive study on the functional and structural properties of variants can provide more sophisticated attributes for the development of prediction methods. This encourages further application and development on the project. We currently also look at the co-evolution analysis of co-localised interface nsSNPs. Those co-localised nsSNPs that cause the same type of disease may be highly associated throughout evolution (co-evolve). nsSNPs^{GD} and nsSNPs^{SC} are expected to show different evolutionary relationships between co-localised nsSNP pairs since these two types of disease-related nsSNPs affect biological system through different mechanisms and exhibit different characteristics. These relationships will be explained in details in future studies.

In future work, we will tackle several issues to further improve this work. First, the human PPI dataset was obtained and integrated from a number of public databases. The database STRING provides both experimental and predicted PPIs. These PPIs were used to extend 3D space of the human 3D PPINs. However, many of these predicted PPIs may not occur in vivo and should be either removed from the network or filtered using stringent criteria, such as that both proteins of interaction pairs involve in the same functional pathway or have resolved crystal structures in the same complex.

The second area we would seek to include is to use a more appropriate statistical method to estimate the level of differences between samples which have large sample sizes. In

this work, the Wilcoxon Mann-Whitney test was used to estimate the level of differences between re-sampled propensities at studied protein regions. The randomly re-sampled propensities were generated using bootstrapping method with 10,000 replicates. However, statistical tests can skew the results when the sample size is large [227]. The differences between most of the compared SNP classes showed statistically significant in our results. A method proposed by Wolfe and Hanley [227] takes into account the level of overlapping confidence intervals between two samples. This may reduce the effect of the sample size.

Lastly, the somatic cancer variant dataset (nsSNPs^{SC}) obtained from the COSMIC database contains both driver and passenger variants. An effective method is required to distinguish drivers from passengers. One approach is to look at only the variants that occur at driver genes. A previous work by Futreal *et al.* [228] listing 291 experimentally identified driver genes provides a reliable resource for studying the mutations of the driver genes. Although mutations found in driver genes are not necessarily driver mutations, by only looking at these mutations we may reduce the statistical bias coming from passenger mutations. Another approach is to use the prediction tools, such as MutationAssessor and MuSiC (mentioned in Chapter 2), to predict mutations which are prone to have functional or structural effect on proteins. A meta-analysis of the predicted results from several different prediction tools would generate a list of candidate mutations which are most likely to be driver mutations. This may bring us insights into the nature of the mutations which play a role in oncogenesis.

Appendix A

Supplementary Data

	Stiff	mean(Size$_{stiff}$) ± SEM	Flexible	mean(Size$_{flexible}$) ± SEM
nsSNPs$^{C}_{surface}$	24,051	73.84±0.74	17,694	53.34±0.54
nsSNPs$^{C}_{interface}$	8,152	25.70±0.29	6,600	21.11±0.25
nsSNPs$^{C}_{core}$	34,797	110.90±2.36	7,170	24.22±0.61
nsSNPs$^{GD}_{surface}$	1,450	91.63±2.09	979	67.37±1.63
nsSNPs$^{GD}_{interface}$	660	32.63±0.88	469	26.91±0.76
nsSNPs$^{GD}_{core}$	1,569	157.14±8.35	389	29.58±2.50
nsSNPs$^{SC}_{surface}$	12,158	74.52±0.76	8,788	53.71±0.55
nsSNPs$^{SC}_{interface}$	4,449	25.95±0.30	3,801	21.30±0.25
nsSNPs$^{SC}_{core}$	17,102	115.93±3.91	3,795	24.75±0.66

Table A.1: Numbers of nsSNPs mapped on secondary structure elements. nsSNPs^{C}: general common nsSNPs; nsSNPs^{GD}: germ-line disease nsSNPs; nsSNPs^{SC}: somatic cancer nsSNPs. The number of nsSNPs from each class and the average size (length in amino acid residues) of the stiff/flexible regions are listed. The standard error of the mean (SEM) is also reported.

	Surface		Interface		Core	
	stiff	flexible	stiff	flexible	stiff	flexible
Drastic Change						
nsSNPs^{C}	5,110	5,075	1,787	1,863	2,028	1,891
nsSNPs^{GD}	485	334	214	175	227	110
nsSNPs^{SC}	2,994	2,716	1,258	1,333	1,394	1,040
Moderate Change						
nsSNPs^{C}	8,642	6,689	3,107	2,482	4,078	2,622
nsSNPs^{GD}	422	296	186	147	230	137
nsSNPs^{SC}	4,937	3,750	1,862	1,801	2,421	1,549

Table A.2: Numbers of nsSNPs mapped on secondary structure elements and amino acid change type.

	Protein Num.	Domain	Surface	Interface	Core	Disordered
Total	2,100		3,707	1,630	1,880	288
A. at functional sites						
nsSNPs^{C}	229	259	141	54	64	11
nsSNPs^{GD}	49	61	36	15	10	6
nsSNPs^{SC}	163	167	82	43	42	15
B. sequence-based screening						
nsSNPs^{C}	695	1,131	540	247	344	54
nsSNPs^{GD}	112	200	108	50	42	14
nsSNPs^{SC}	456	605	304	138	163	40
C. screening in 3D space						
nsSNPs^{C}	915		969	374	572	
nsSNPs^{GD}	140		175	62	83	
nsSNPs^{SC}	700		525	238	294	

Table A.3: Numbers of nsSNPs close to functional sites.

	Protein Num.	Domain	Surface	Interface	Core	Disordered
Total	4,369		4,333	2,086	4,153	18,088
A. at PTM sites						
nsSNPs^{C}	814	391	162	72	157	715
nsSNPs^{GD}	60	55	26	15	14	21
nsSNPs^{SC}	440	258	94	68	96	320
B. sequence-based screening						
nsSNPs^{C}	1,951	1,832	730	309	793	3,288
nsSNPs^{GD}	123	202	95	53	54	72
nsSNPs^{SC}	1,202	1,063	403	242	418	1,431
C. screening in 3D space						
nsSNPs^{C}	1,157		1,392	543	654	
nsSNPs^{GD}	137		169	69	86	
nsSNPs^{SC}	858		864	426	405	

Table A.4: Numbers of nsSNPs close to post-translational modification sites.

	Surface	Interface	Core	Disordered
Bone	25	4	23	6
Cancer	142	48	74	73
Cardiovascular	57	5	56	41
Connective_tissue	30	7	41	10
Dermatological	45	8	106	43
Developmental	63	14	42	32
Ear_Nose_Throat	27	4	8	12
Endocrine	120	8	101	44
Gastrointestinal	22	2	13	10
Hematological	529	10	146	60
Immunological	84	5	73	27
Metabolic	330	15	275	42
multiple	134	17	119	57
Muscular	119	6	91	39
Neurological	148	18	137	43
Nutritional	9	2	4	5
Ophthamological	81	9	72	28
Psychiatric	10	1	8	4
Renal	32	1	33	19
Respiratory	4	1	5	0
Skeletal	82	5	58	32

Table A.5: Numbers of nsSNPs^{GD} by disease types.

	Surface	Interface	Core	Disordered
Adrenal gland	8	4	6	7
Autonomic ganglia	87	27	78	82
Biliary tract	44	36	42	55
Bone	31	8	15	24
Breast	1,156	487	1,238	1,516
Central nervous system	663	290	641	784
Cervix	221	68	191	309
Endometrium	2,639	1,076	2,649	2,912
Eye	8	11	1	1
Haematopoietic and lymphoid tissue	981	539	990	1,138
Kidney	1,459	471	1,424	1,606
Large intestine	7,194	2,746	7,399	7,840
Liver	134	66	109	140
Lung	3,181	1,448	2,671	3,646
Meninges	45	15	47	44
Oesophagus	150	54	134	186
Ovary	1,414	548	1,468	1,650
Pancreas	446	195	463	459
Prostate	827	328	884	925
Skin	1,056	419	944	1,244
Small intestine	3	10	1	4
Soft tissue	25	36	16	46
Stomach	137	86	124	173
Testis	2	13	1	1
Thyroid	40	42	68	65
Upper aerodigestive tract	191	107	206	229
Urinary tract	491	222	493	678

Table A.6: Numbers of nsSNPs^{SC} by cancer types.

Appendix B

The Publication

EXPERT REVIEWS

Protein–protein interaction networks studies and importance of 3D structure knowledge

Expert Rev. Proteomics 10(6), 511–520 (2013)

Hui-Chun Lu, Arianna Fornili and Franca Fraternali*

Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London SE1 1UL, UK
**Author for correspondence: franca.fraternali@kcl.ac.uk*

Protein–protein interaction networks (PPINs) are a powerful tool to study biological processes in living cells. In this review, we present the progress of PPIN studies from abstract to more detailed representations. We will focus on 3D interactome networks, which offer detailed information at the atomic level. This information can be exploited in understanding not only the underlying cellular mechanisms, but also how human variants and disease-causing mutations affect protein functions and complexes' stability. Recent studies have used structural information on PPINs to also understand the molecular mechanisms of binding partner selection. We will address the challenges in generating 3D PPINs due to the restricted number of solved protein structures. Finally, some of the current use of 3D PPINs will be discussed, highlighting their contribution to the studies in genotype–phenotype relationships and in the optimization of targeted studies to design novel chemical compounds for medical treatments.

KEYWORDS: 3D protein–protein interaction networks • disease-causing mutations • drug design • nsSNPs • protein complexes interface regions • protein–protein interaction networks

Most biological processes in a living cell, such as transcription regulation, signal transduction and cell motility, are mediated by protein–protein interactions (PPIs). Network representations effectively address the complexity of PPIs in biological systems. Indeed, networks provide a highly compact and comprehensive view of binary relationships, in which nodes represent proteins/genes and edges indicate functional association or physical interactions between protein/gene pairs (FIGURE 1A). Many studies used network analysis to report proteins/genes involved in a particular disease or function [1,2] and to complement large-scale siRNA screenings [3,4]. The network topological properties of genes, including the degree of interactions, the clustering coefficient and the betweenness, which are the measurements of connectivity, interconnectivity and centrality, respectively, are often used to characterize topological features of a network [5,6] and provide useful insights: for example, cancer proteins were found to have distinguishable topological features to the other proteins. These are generally found to be hub proteins with relatively high centrality [7,8]. Network representations can also be used to study the associations between objects. In the study of Goh *et al.* [9], a human diseasome bipartite network was constructed to model associations between genes and diseases. A link between a disease and a disease gene indicates that mutations in that gene are resulting in the specific disease. Many diseases were found to share a common genetic origin. This was an interesting finding which suggested that diseases may not be as independent of each other as we know from the traditional clinical assessment.

Apart from disease-related studies, PPINs are also often used in functional studies to assign putative functions to newly discovered genes [10,11] using algorithms based on the 'Guilt-by-Association' principle (FIGURE 1B). This has been particularly explored for plant-specific proteins [12–14] as the majority of these protein functions remain unknown, and yet it is critical to understand their biological relevance and the involving biological processes,

 10.1586/14789450.2013.856764 ISSN 1478-9450

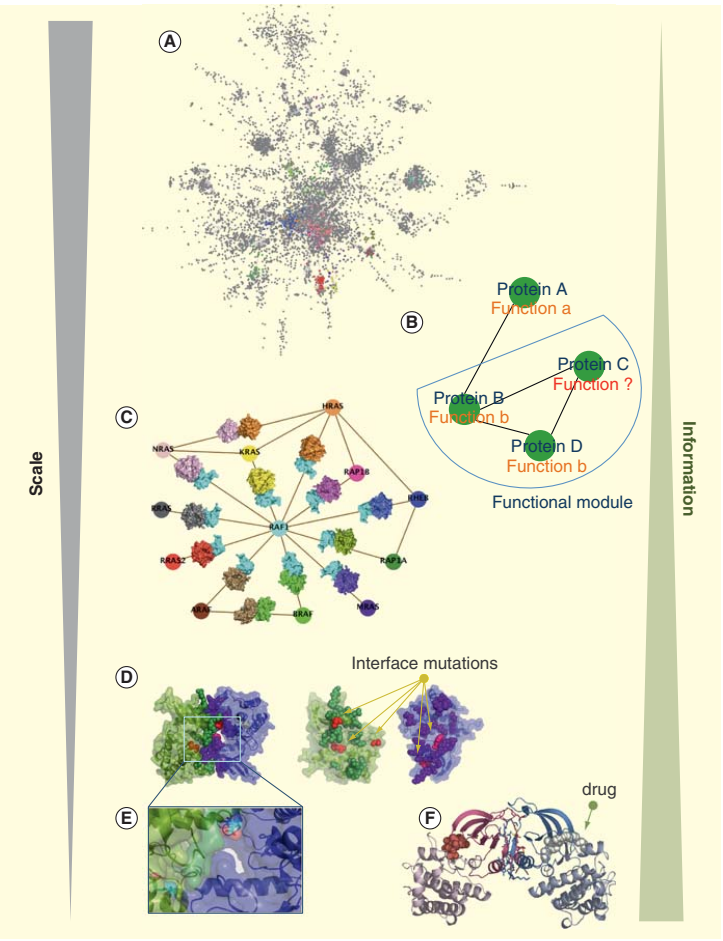

Figure 1. The scale of the biological system versus the amount of information a protein–protein interaction network contains. (A) The human protein–protein interaction network. **(B)** 'Guilt-by-Association' principle in predicting protein functions. **(C)** 3D RAF1 subnetwork containing information from system level to atomic level. The nodes represent proteins, whereas the edges are annotated with protein complexes. **(D)** The structure of protein complex. The interface regions are shown with shape sphere and mapped with disease-causing nsSNVs (colored in red). **(E)** The atomic-level view of interface with nsSNVs. **(F)** A drug inhibits the binding pocket.

including growth control and genotype–phenotype relationships for the variety of plants.

Another approach is based on integrating genomic information and/or PPIN with the knowledge of functional pathways [15], which can be retrieved from databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [16]. These mapping has been particularly exploited for cancer studies. Genomic perturbations attributed to diseases can be mapped on biochemical pathways so as to obtain a pathway-level understanding at distinct disease states. It has been demonstrated that when cancer patients harbor genomic alterations or aberrant expression of different genes, these participate in a common pathway or have a similar effect in altering the pathway [17,18].

These studies supported by PPINs have brought novel insight into cellular functional modules and the association between genes and diseases. By integrating PPINs with the atomic-level information (Figure 1C), one can understand more precise details on the mechanisms regulating how proteins specify their functions and how disease-causing mutations disrupt the biologically functional systems.

Pioneering studies in structure-based PPINs were done by Aloy and Russell [19] who looked into atomic details of protein interaction pairs and proposed that homologous protein pairs may interact in the same way, using the same binding interfaces. In the effort of bridging the gap between large-scale PPI determination and structural data, hybrid approaches to structure determination of macromolecular complexes have been proposed [20–22]. Integration of structural data at different resolution and reliability has been successfully used to reconstruct hybrid assembly structures that can be informative for further PPI validation studies or in designing new tailored structural investigations. These 3D network studies were initially done on the model organism yeast. Recently, in reason of the increasing number of available structures and interaction data, 3D PPINs of other organisms have also become available [23,24], including human, mouse, drosophila and bacteria.

Recent studies [25–29] integrated protein structural information with PPINs to give the ability to implement large-scale studies on the association between cellular mechanism and protein complexes. One of the major interests in exploiting these studies is to investigate how disease-related mutations may disrupt protein functions and ultimately affect the function of biological systems. Mutations can be classified as loss of function, gain of function or neutral according to their effect on protein function. These effects can be mediated by alterations of the protein stability induced by the mutation [30,31]. For example, the B-RAF kinase is widely mutated in cancer and the V600E mutant, recently observed also in patients with granulomatous pediatric disease [32], destabilizes the inactive

conformation of the kinase. This gain-of-function mutant keeps B-RAF in the active state and consequently increases the activation of another kinase ERK [32]. Mutations can affect protein function also by modifying the affinity of the protein for its partners. For example, mutations derived from Glioblastoma patients have been recently shown to destabilize the complexes of the proteins involved in the disease pathogenesis, mainly through a decrease of the electrostatic contributions to the binding energy [27]. Drastic amino acid changes at the protein interface, where a protein is in physical contact with another protein, can significantly change the binding energy of the interaction. In particular, the occurrence of mutations at the interface hot spot residues, which contribute the most to the binding energy [33,34], are most likely to have impact on the interaction, so that the proteins would either loose interactions with the partner proteins or gain interactions with new binding proteins. Additionally, protein function can be altered by mutations occurring at allosteric sites. Indeed, allosteric mutations can disrupt or promote the binding of allosteric modulators, affect the communication pathways between the allosteric and orthosteric sites and modify the relative proportion of inactive/active conformations [35,36]. For example, different cancer-related mutations in kinases have been shown to involve a shifting in the relative population of the inactive/active states [37,38].

Besides providing insight into the impact on the complex functionality of disease-causing mutations, 3D PPINs can be used for large-scale screenings of drug targets to compensate experimental drug compound screening methods [39]. A number of approaches have been developed in the recent years to exploit PPINs in drug discovery, as the cellular network and the surrounding environment are essential part of the process of efficient drug targeting and delivering. By implementing large-scale screening over protein molecular properties, one could identify new target proteins and potential binding sites of drug compounds. In particular, PPINs can be used to identify PPI inhibitors. Targeting PPI is still one of the most challenging task in drug design, owing to the significant differences between interfaces in PPIs and small molecule-binding sites. However, specific properties embedded in PPI interfaces and needed for partners recognition can be exploited to identify drug compound targets [40–44]. Indeed, there are increasing examples of targeting PPIs (Figure 1F) with drugs that can bind to transient and dynamic pockets in orthosteric or allosteric sites (for a review on the subject, see Engin *et al.* [45] and references therein). In targeting PPIs, people are also exploiting protein interface motifs to identify potential off-target drugs [46]. Databases such as Interactome3D [23] and INstruct [24] provide PPINs annotated with protein complexes and are essential to large-scale screening approaches targeting protein interfaces in drug design [47].

Another interesting use of targeting PPIN, pathways and drug action mechanisms is the identification of proteins that induce drug side effects [48]. Drug side effects are often the most undesirable outcomes from medical treatment, frequently caused by the binding between drug compounds and off-target proteins. Adverse drug reaction was reported to be one of the major causes of mortality and morbidity over the last decades [49]. Polypharmacology approaches have become popular in complementing the classical 'one-drug one-target' paradigm [50].

Still, the bottleneck of systematic screening of binding pockets for drug compounds in silico lays on the limited availability of experimental structures. Homology modeling can be used in generating 3D protein models to compensate this limitation [51,52]. However, high-quality structures are essential for binding pocket detection. Model refinement procedures can help in obtaining a more realistic structure for these drug target-binding studies [53]. However, additional challenges in protein complex prediction are in that often proteins are subject to conformational changes to attain specific binding modes. A special case is when these functional states are induced by allosteric sites signaling, generally not easy to observe experimentally [54]. Thus, conformational change perturbations are usually not taken into account in protein complexes modeling procedures.

In the following sections, we discuss the availability and quality of PPI data sets, as well as the current state of high-throughput experimental methods for PPIs detection since they are fundamental to build a 3D PPIN. We particularly focus on recent applications of 3D PPIN, highlighting strengths and discussing limitations related to the availability of structural data for human proteins. Finally, we briefly comment on how 3D PPIN could contribute to the design of novel-targeted therapies, particularly useful to the advancement of personalized medicine.

Protein–protein interaction data

High-throughput experimental methods have given the possibility to build PPINs of entire organisms, which in some cases (e.g., yeast) [55] are deemed close to complete. They include detection of direct interactions by yeast two-hybrid assays and detection of protein complexes by affinity purification-mass spectrometry (AP-MS). Literature curation and annotation are another useful source for PPI data sets, often extracting information obtained from small-scale experiments, such as Fluorescence Resonance Energy Transfer or other biophysical investigations. However, these collected data sets are usually biased toward the proteins that have been most extensively studied and are not large scale, due to constraints in the detection methods. In 2008, it was estimated that about 650,000 PPIs should occur in humans [56] and so far about one-tenth of the estimated human interactions have been observed experimentally [57]. Publicly available databases, such as HPRD [58], BioGRID [59], DIP [60], IntAct [61], MINT [62] and STRING [63], provide platforms to access PPIs curated data sets.

Although the high-throughput experiment techniques are progressing to obtain complete pictures of biological systems, low reproducibility of the data has raised concern about the data quality. Braun [64] pointed out that the overlap of yeast PPI data sets derived from AP-MS experiments between two

labs could be as low as 20%. This may be ascribed to different reasons including the absence of the same standardized experiment protocols and biased sampling. Varjosalo and colleagues [65] demonstrated that high-throughput PPI experiments are highly reproducible when performed by two different labs if the protocols with the same standardized workflows are used. Moreover, to diminish bias sampling, they used 32 human kinases as bait proteins with a different domain composition, expressed in different tissues and involved in different biological processes. Analogously, Havugimana and colleagues [66] generated a pipeline with stringent experiment procedures and applying computational methods to detect high-abundance components and identify functionally unrelated protein pairs. Proteomic profiles were used to assess the abundance, reducing the number of false-positive interactions from protein pairs that *in vivo* are not expressed at the same time and cellular space. Scientists in the AP-MS field are developing experimental approaches to mitigate some of these inefficiencies, using, for example, replicated and control experiments and relative quantification to enhance sensitivity and/or by developing confidence scores to select specific PPIs [67–69]. Apart from experiment protocols, many studies also suggested the need of data standardization [70,71] and validation [67,72]. The International Molecular Exchange consortium provides the controlled vocabularies and standardized data formats that have been adopted by major databases [73,74]. Statistical methods [75] and structural information are also suggested for the validation of PPIs before they are deposited to the databases.

Additionally, a number of computational methods have been developed to compensate the experimental methods and expand the space of PPINs based on strategies such as co-evolution [76] and homology modeling [77] (for a recent review on computational prediction methods, see Mosca *et al.* [71]). However, high-confidence PPI data with experimental evidence are fundamental to build 3D PPINs, which could carry out more robust studies in disease-related mutations and drug target identification.

The construction of 3D PPINs

As previously mentioned, PPIN is a useful tool in identifying disease or functional relationships between proteins. Yet, system-level representations of biological processes provide very limited information on answering crucial questions, such as how a protein recognizes its partner proteins, or which region of its surface binds to its partner proteins. It requires atomic-level information to understand binding mechanisms. However, mapping structural information onto networks remains a challenge due to the gap between the number of known proteins and the number of solved protein structures and some types of proteins are under-represented in structure databases such as membrane proteins. So far, the Protein Data Bank [78] stores roughly around 5,000 human protein structures with many of them containing only partial structures.

To tackle this problem, earlier work in developing 3D molecular interaction networks, including iPfam [79] and 3did [80], analyzed the structures of protein complexes at domain level. The domain-based interactions are supported by both inter- and intraspecies co-crystal structures, and they include interactions between domains belonging both to the same and to different proteins. 3did covers more than 4,000 distinct domains that are about one-third of the total number of Pfam domains [81]. Importantly, domains are the basic evolutionary and functional units of proteins. Proteins with domains in a common superfamily are considered more likely to be evolutionarily related [82]. By looking at the molecular details of protein domain interactions, one could identify the domains that are functionally important in mediating PPIs.

To increase the coverage of structures in PPINs, the two recent databases INstruct [24] and Interactome3D [23] implement two different approaches both based on the use of homologous structures. One should be aware that PPI predictions using homologous structures can be of different nature. One is to use homologous protein structures of a protein pairs to predict the possibility of interaction, which is not detected from experimental methods. The other is to predict structure complex of a protein pair, which is known to interact from experimental methods and therefore trying to enrich with structural information the available large-scale screens. The following discussions are based on the second strategy to predict the protein complexes. The pipeline of INstruct to generate a 3D PPIN starts from binary interaction data sets from different publicly available databases. Each interacting pair is then annotated with the corresponding co-crystal structure if available or with co-crystal structures of homologous proteins. It should be noted that the resulting structural annotation of protein pairs with homologous co-crystals is only approximate. The database provides 3D PPIN data of human and six other most studied model organisms, where human 3D PPIN contains 6,585 interactions between 3,627 proteins. A different strategy was used for Interactome3D, where the structural coverage of human PPIN was increased by modeling interacting pairs with missing structural data using Modeller [83]. This provides a more precise representation of interface regions of interacting protein pairs. Interface residues are identified by calculating the distance of residues from protein pairs. The database provides human 3D PPIN, which contains 6,473 interactions between 4,239 proteins.

One may also use predicted protein structures obtained from reliable resources to compensate the limitation of solved protein structures. A recent project Genome3D [84] integrates UK-based structural resources, including Gene3D [85], FUGUE [86] and four other structural prediction resources [14,87–89]. The aims of this project are to provide biologists with a platform to compare the predictions from those resources that were developed with different algorithms, and to choose the prediction outcomes that are more reliable. Those predicted structures could help to expand the size of 3D PPINs.

To identify or even predict the proteins that can be acting together, one could use available structural information. A study by Kar *et al.* [90], for instance, looked at the structural features of cancer proteins. Cancer proteins are well known to

be involved in biological processes related to, for example, DNA repair and cell growth. In this study, 10 functional pathways were selected according to the Cancer Cell Map [201]. Each protein pair in a given functional pathway is annotated with structural information obtained by running PRISM [91], a software to explore the known protein–protein interface-binding modes and predict analogous cases. PRISM can predict the protein interaction by searching interface with similar backbone geometry in the interface library. In this way, co-crystal structures are not strictly required to build the PPIN of the pathway. This method is beneficial for smaller scale studies with interests in proteins in particular functional pathways and save computational time to construct a 3D PPIN of an entire proteome.

To summarize, in this section, we have reviewed some of the recent approaches to generate 3D PPINs. As it will be shown in the following sections, building increasingly complete 3D PPINs is essential to interpret biological systems, provide insight into complex cellular mechanisms and rationalize genotype–phenotype relationships.

Protein interaction interfaces

The ability of proteins to recognize and bind their partners is essential to biological processes. Protein interfaces, where proteins have physical contact with their partner proteins, are believed to embed crucial properties that mediate PPIs. Both experiments and computational analyses have shown that protein interfaces mediate PPIs through specific molecular properties, including sequence motifs [46,92], backbone geometry [93], residue types [94], interface hot spots [33,95] and correlated changes in the two interfaces.

Each interface of a protein is composed of several discontinuous patches. Two estimates have been commonly used to define the interface regions. The first approach is to calculate the distances between residue pairs from two proteins in a co-crystal structure [23,43]. Two residues are considered as interacting if their distance is within a predefined threshold. This value may depend on the group of residue atoms that is used to calculate the inter-residue distance. Typical threshold values are 4–5 Å on the distance between any pair of atoms from the two residues [80,96] or 9 Å on the distance between Cα or Cβ atoms [43]. The sum of atomic van der Waals radii + 0.5 Å is another frequently used distance threshold [97]. More sophisticated criteria use different thresholds for different types of interactions (e.g., hydrogen bonds, salt bridges and van der Waals interactions) [23]. The second approach compares the solvent accessible surface area (SASA) of the protein complex with that of the single components to evaluate the area buried upon complex formation (interface area). For example, residues can be considered as part of the interface if their total or side chain buried solvent accessible surface area is larger than 0–1 $Å^2$ [98,99]. The calculation of the buried area can be implemented in automated approaches such as POPSCOMP [100]. Precompiled values for protein complexes are also available from databases such as 3did and PIBASE [101]. For protein pairs where the structure of the single proteins is available but not that of the complex, molecular docking methods, such as FiberDock [102], can be used to predict the interface regions.

In one of the leading studies on biologic structural networks, Kim and colleagues [103], classified hub proteins into two groups according to the number of their interfaces: multi-interface hubs and singlish-interface hubs. The two groups were shown to have different evolutionary properties. In particular, only multi-interface hubs turned out to be significantly more essential and slow evolving compared with the average. Since early studies considered all hub proteins to be essential [104,105], this shows that the integration of sequential and structural information on protein interfaces can increase the precision in identifying functionally important proteins within biological systems.

The importance of interfaces for protein functions and biological processes was further confirmed in recent studies by looking at the occurrences of disease-associated mutations [26,106], where the interface regions were found to be enriched in disease-causing mutations. This implies that a residue change at these region is more likely to disrupt protein functions and lead to diseases. In a follow-up study by the same group [107], further annotations were given to the mutations mapped on 3D PPIN. Dominant truncating disease mutations were found to have different pattern to other classes of mutations with no preference occurring at interface regions. Moreover, recessive mutations co-localized on the same interface showed the tendency to cause the same disease.

Increasing the level of detail in the description of protein interfaces further highlights the importance of structural properties in determining the protein function. For example, promiscuous binding sites, which are essential for hub proteins to interact with many different partners, can be identified by mapping interactions with multiple partners on the protein surface. Promiscuous sites have been shown to possess specific properties in terms of amino acid composition [108,109], solvent accessibility [110], packing [97] and conformational flexibility [111,112], mainly related with their increased capacity to adapt to different partners. The biological relevance of promiscuous residues is further confirmed by a recent study from our laboratory [111], showing that they are less enriched in non-synonymous single-nucleotide variations (nsSNVs). This finding suggests that residues in promiscuous positions have a reduced tolerance to genetic variations, related to the necessity to preserve their binding polyvalence.

Although key challenges remain in performing interactome-scale studies on protein interface regions due to the low availability of protein structures, computational approaches may help in overcoming this limitation. Gao and Skolnick [93] found that many protein complexes have similar interfaces even if the overall structure of the single components is different. Thus, they argued that even though there are only a small fraction of proteins with solved structures, the protein interface library is close to complete. This is the fundamental idea

behind protein-binding pocket searches [113,114] and PPI predictions that use interface structure similarity scoring [77,91] in order to increase the structural coverage of 3D PPIN.

How 3D PPINs contribute to biology and biomedical sciences

Advances in genome-sequencing techniques and large-scale genome-sequencing projects, including the 1000 Genomes Project [115] and the International HapMap Project [116], are boosting the amount of available gene variation data. Databases, including dbSNP [117], OMIM [118], COSMIC [119], HGMD [120] and Exome Variant Server [202], provide online interfaces to easily access gene variation data sets and serve with different purposes. The challenge is now to develop methods and tools to extract useful information from this increasing amount of data. In particular, it will be essential to understand what information those mutations are carrying, how we can use those gene variation data to unravel the underlying cellular mechanisms and how disease-causing mutations lead to diseases [121,27]. To answer these questions, atomic-level information of protein complexes is crucial to provide the biological features of disease-related proteins and disease-causing mutations. Moreover, by implementing large-scale studies with 3D PPINs, which contain information from system level to atomic level, one may find explanations for the effects of these mutations. For example, Kar *et al.* [90] implemented human structural protein interface network, in which the edges represent binding interfaces between protein pairs obtained from either known or predicted structures from PRISM. Their results show that cancer proteins tend to be hubs in the network. The mutations occurring at cancer protein interfaces, disrupting protein bindings and causing loss of protein functions, have greater impact on biological systems. The binding interfaces of cancer proteins were also shown to have specific properties; they are significantly smaller, more planar, less compact and less hydrophobic than interfaces in noncancer proteins. These specific features of cancer protein interfaces may be used for the identification of new targets and drug candidates in cancer therapies. The results demonstrated how 3D PPINs can enable more comprehensive studies in biological systems with informative outcomes. Besides, 3D PPINs can be effectively used in high-throughput screening for drug targets. Indeed, PPIs are promising druggable targets since their selective inhibition [122] can be used to regulate particular functions in biological systems. 3D PPINs, by representing in a synthetic and comprehensive way the associations between a target protein and its partner proteins, are an invaluable tool to understand the possible effects of inhibiting the binding interfaces of the target protein. All these examples show how 3D PPINs can give an essential contribution to Biomedical Sciences.

Expert commentary

The studies of PPINs have progressed from system-level representations of biological systems to more detailed representations annotated with atomic information. The integration of data from the currently available biological databases has proven to be essential to carry out comprehensive studies on cellular mechanisms and the causes of their disruption. 3D PPINs analysis has been used to study the role of disease-causing mutations. So far, despite the general agreement on the propensity of disease-related mutations for protein interface regions, the general characteristics of these mutations and how they affect biological functions remain challenging. Still, 3D PPINs analysis is very important to unravel the features of these mutations and their impact on protein functions. In particular, the identification of properties specific to pathogenic mutants is crucial to develop methods for the prediction of disease-related mutations. Besides, 3D PPINs analysis can also help biologists to effectively search for possible targets for disease treatment as PPIs are ideal drug targets [122] to regulate biological functions, and the structural information in 3D networks provides the guidance for the design of drug compounds in the early stage development. 3D PPINs provide fundamental materials for the screening of off-target PPIs. The use of these preliminary investigation strategies could effectively reduce time and cost in drug development. The increased understanding of the pathogenic mechanisms triggered by disease mutations, and of the activity of drug compounds in the cell, combined with personal genome sequencing profiles, will promote the development and delivery of more effective personalized clinical treatments in the foreseeable future.

Five-year view

The studies using 3D PPINs are a relatively new research field. The bottleneck in the generation of complete 3D PPINs lays in the limitation of available experimental data on genome sequences and protein structures. With the rapid progressing of experimental technologies and bioinformatics approaches, more biological data will become available to provide a more complete view of biological systems. In the coming years, the importance of protein-binding mechanisms and the general characteristics of disease mutations will be better understood. Therefore, it will be possible to develop more sensitive predictors of disease-causing mutations, resulting in further progress toward effective personalized medical treatments.

Financial & competing interests disclosure

H-C Lu acknowledges support from King's College London (London, UK). A Fornili and F Fraternali acknowledge support from the British Heart Foundation. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties

No writing assistance was utilized in the production of this manuscript.

Key issues

- Currently, the number of solved protein structures is relatively small when compared with the number of known proteins for each species. Although homologous structure modeling could narrow the gap, a close-to-complete map of biological system requires a higher number of available structures.
- Modeling protein complexes remains challenging.
- Some types of proteins are significantly under-represented in protein structure databases such as membrane proteins. These proteins are also determinant for cellular signaling and drug target studies.
- It is now timely to develop methods and tools to extract useful information from the large amount of data generated by large-scale genome, metabolome, transcriptome and proteome projects.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1 Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA* 105(11), 4323–4328 (2008).

2 Glaab E, Baudot A, Krasnogor N, Valencia A. Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics* 11, 597 (2010).

3 Carlin LM, Evans R, Milewicz H *et al.* A targeted siRNA screen identifies regulators of Cdc42 activity at the natural killer cell immunological synapse. *Sci. Signal.* 4(201), ra81 (2011).

4 Bakal C, Linding R, Llense F *et al.* Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science* 322(5900), 453–456 (2008).

5 Yamada T, Bork P. Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat. Rev. Mol. Cell Biol.* 10(11), 791–803 (2009).

6 Gursoy A, Keskin O, Nussinov R. Topological properties of protein interaction networks from a structural perspective. *Biochem. Soc. Trans.* 36(Pt. 6), 1398–1403 (2008).

7 Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22(18), 2291–2297 (2006).

8 Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD. Low duplicability and network fragility of cancer genes. *Trends. Genet.* 24(9), 427–430 (2008).

9 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc. Natl Acad. Sci. USA* 104(21), 8685–8690 (2007).

10 Karaoz U, Murali TM, Letovsky S *et al.* Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA* 101(9), 2888–2893 (2004).

11 Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21(Suppl. 1), i302–i310 (2005).

12 Benfey PN, Mitchell-Olds T. From genotype to phenotype: systems biology meets natural variation. *Science* 320(5875), 495–497 (2008).

13 Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an arabidopsis interactome map. *Science* 333(6042), 601–607 (2011).

14 Lee I, Seo YS, Coltrane D *et al.* Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc. Natl Acad. Sci. USA* 108(45), 18548–18553 (2011).

15 Kuzu G, Keskin O, Gursoy A, Nussinov R. Constructing structural networks of signaling pathways on the proteome scale. *Curr. Opin. Struct. Biol.* 22(3), 367–377 (2012).

16 Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40(Database issue): D109–D114 (2012).

17 Feng J, Kim ST, Liu W *et al.* An integrated analysis of germline and somatic, genetic and epigenetic alterations at 9p21.3 in glioblastoma. *Cancer* 118(1), 232–240 (2012).

18 Mathew JP, Taylor BS, Bader GD *et al.* From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput. Biol.* 3(2), e12 (2007).

19 Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* 99(9), 5896–5901 (2002).

20 Aloy P, Böttcher B, Ceulemans H *et al.* Structure-based assembly of protein complexes in yeast. *Science* 303(5666), 2026–2029 (2004).

21 Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. *Nature* 450(7172), 973–982 (2007).

22 Alber F, Dokudovskaya S, Veenhoff LM *et al.* Determining the architectures of macromolecular assemblies. *Nature* 450(7170), 683–694 (2007).

23 Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. *Nat. Methods.* 10(1), 47–53 (2012).

•• **Currently, one of publicly available databases that provide most reliable and reliable 3D protein–protein interaction networks data in different species.**

24 Meyer MJ, Das J, Wang X, Yu H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29(12), 1577–1579 (2013).

25 Bockler B, Bateman A. Protein interactions in human genetic diseases. *Genome Biol.* 9(1), R9 (2008).

26 Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30(2), 159–164 (2012).

• **A pioneering study mapping human mutations onto protein structures in the human 3D protein–protein interaction networks.**

27 Nishi H, Tyagi M, Teng S *et al.* Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS ONE* 8(6), e66273 (2013).

28 Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA. Systematic computational prediction of protein interaction networks. *Phys. Biol.* 8(3), 035008 (2011).

29 Hooda Y, Kim PM. Computational structural analysis of protein interactions and networks. *Proteomics* 12(10), 1697–1705 (2012).

30 Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353(2), 459–473 (2005).

31 Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* 449(3), 581–594 (2013).

32 Satoh T, Smith A, Sarde A *et al.* B-RAF mutant alleles associated with Langerhans cell histiocytosis, a granulomatous pediatric disease. *PLoS ONE* 7(4), e33891 (2012).

33 Keskin O, Ma B, Nussinov R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* 345(5), 1281–1294 (2005).

34 Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* 280(1), 1–9 (1998).

35 Liu J, Nussinov R. Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design. *Proc. Natl Acad. Sci. USA* 105(3), 901–906 (2008).

36 Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. *Cell* 153(2), 293–305 (2013).

37 Shan Y, Eastwood MP, Zhang X *et al.* Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* 149(4), 860–870 (2012).

38 Azam M, Seeliger MA, Gray NS, Kuriyan J, Daley GQ. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat. Struct. Mol. Biol.* 15(10), 1109–1118 (2008).

39 Blundell TL, Sibanda BL, Montalvão RW *et al.* Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361 (1467), 413–423 (2006).

•• **An interesting review on finding protein-binding pocket as potential drug target assisted by structural information and computational methods.**

40 Azzarito V, Long K, Murphy NS, Wilson AJ. Inhibition of -helix-mediated protein-protein interactions using designed molecules. *Nat. Chem.* 5, 161–173 (2013).

41 Kastritis PL, Bonvin AM. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface* 10(79), 20120835 (2013).

42 Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450(7172), 1001–1009 (2007).

43 Winter C, Henschel A, Tuukkanen A, Schroeder M. Protein interactions in 3D: from interface evolution to drug discovery. *J. Struct. Biol.* 179(3), 347–358 (2012).

44 Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* 108(4), 1225–1244 (2008).

45 Engin HB, Gursoy A, Nussinov R, Keskin O. Network-based strategies can help mono- and poly-pharmacology drug discovery: a systems biology view. *Curr. Pharm. Des.* doi:10.2174/ 13816128113199990066 (2013) (Epub ahead of print).

46 Engin HB, Keskin O, Nussinov R, Gursoy A. A strategy based on protein-protein interface motifs may help in identifying drug off-targets. *J. Chem. Inf. Model.* 52(8), 2273–2286 (2012).

47 Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* 138(3), 333–408 (2013).

48 Kuhn M, Al Banchaabouchi M, Campillos M, *et al.* Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.* 9, 663 (2013).

49 Wu TY, Jen MH, Bottle A *et al.* Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. *J. R. Soc. Med.* 103(6), 239–250 (2010).

50 Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug-target network. *Nat. Biotechnol.* 25(10), 1119–1126 (2007).

51 Zhang Q, Petrey D, Norel R, Honig B. Protein interface conservation across structure space. *Proc. Natl Acad. Sci. USA* 107(24), 10896–10901 (2010).

52 Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl Acad. Sci. USA* 109(24), 9438–9441 (2012).

53 Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. *Structure* 10(3), 435–440 (2001).

54 Schwede T. Protein modeling: what happened to the "protein structure gap"? *Structure* 21(9), 1531–1540 (2013).

• **A comprehensive review introduces currently available tools for protein structure modeling and prediction and the challenges in the field.**

55 Krogan NJ, Cagney G, Yu H *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440(7084), 637–643 (2006).

56 Stumpf MPH, Thorne T, de Silva E *et al.* Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA* 105(19), 6959–6964 (2008).

57 Lehne B, Schlitt T. Protein-protein interaction databases: keeping up with growing interactomes. *Hum. Genomics* 3(3), 291–297 (2009).

58 Keshava Prasad TS, Goel R, Kandasamy K *et al.* Human protein reference database-2009 update. *Nucleic Acids Res.* 37, D767–D772 (2009).

59 Chatr Aryamontri A, Breitkreutz BJ, Heinicke S *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–D823 (2013).

60 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451 (2004).

61 Kerrien S, Aranda B, Breuza L *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846 (2012).

62 Ceol A, Chatr Aryamontri A, Licata L *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 38, D532–D539 (2009).

63 Franceschini A, Szklarczyk D, Frankild S *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815 (2013).

64 Braun P. Reproducibility restored – on toward the human interactome. *Nat. Methods* 10(4), 301–303 (2013).

65 Varjosalo M, Sacco R, Stukalov A *et al.* Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat. Methods* 10(4), 307–314 (2013).

66 Havugimana PC, Hart GT, Nepusz T *et al.* A census of human soluble protein complexes. *Cell* 150(5), 1068–1081 (2012).

67 Braun P, Tasan M, Dreze M *et al.* An experimentally derived confidence score for

binary protein-protein interactions. *Nat. Methods* 6(1), 91–97 (2009).

68 Jain S, Bader GD. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* 11, 562 (2010).

69 Dazard JE, Saha S, Ewing RM. ROCS: A reproducibility index and confidence score for interaction proteomics studies. *BMC Bioinformatics* 13, 128(2012).

70 Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology – what's the connection? *Nat. Biotechnol.* 26(1), 69–72 (2008).

71 Mosca R, Pons T, Céol A, Valencia A, Aloy P. Towards a detailed atlas of protein-protein interactions. *Curr. Opin. Struct. Biol.* doi:10.1016/j.sbi.2013.07.005 (2013) (Epub ahead of print).

72 von Mering C, Krause R, Snel B *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417(6887), 399–403 (2002).

73 Orchard S, Kerrien S, Abbani S *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Meth.* 9(4), 345–350 (2012).

74 Orchard S. Molecular interaction databases. *Proteomics* 12(10), 1656–1662 (2012).

75 Gavin AC, Aloy P, Grandi P *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440(7084), 631–636 (2006).

76 de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14(4), 249–261 (2013).

77 Zhang QC, Petrey D, Deng L *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature.* 490(7421), 556–560 (2012).

78 Berman HM. The protein data bank. *Nucleic Acids Res.* 28(1), 235–242 (2000).

79 Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics.* 21(3), 410–412 (2005).

80 Stein A, Céol A, Aloy P. 3did: identification and classification of domain based interactions of known three-dimensional structure. *Nucleic Acids Res.* 39(Database issue), D718–D723 (2011).

81 Punta M, Coggill P, Eberhardt R *et al.* The Pfam protein families database. *Nucleic Acids Res.* 40(Database issue), D290–D301 (2012).

82 Vogel C, Teichmann S, Pereiraleal J. The relationship between domain duplication and recombination. *J Mol Biol.* 346(1), 355–365 (2005).

83 Eswar N, Webb B, Marti-Renom MA *et al.* Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* Chapter 2, Unit 2.9 (2007).

84 Lewis TE, Sillitoe I, Andreeva A *et al.* Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res.* 41(Database issue), D499–D507 (2013).

85 Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.* 38(Database issue), D296–D300 (2010).

86 Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol.* 310(1), 243–257 (2001).

87 Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30(1), 268–272 (2002).

88 Bennett-Lovsey RM, Herbert AD, Sternberg MJE, Kelley LA. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins* 70(3), 611–625 (2008).

89 Buchan DWA, Ward SM, Lobley AE, Nugent TCO, Bryson K, Jones DT. Protein annotation and modelling servers at University College London. *Nucleic Acids Res.* 38(Web Server issue), W563–W568 (2010).

90 Kar G, Gursoy A, Keskin O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput. Biol.* 5(12), e1000601 (2009).

91 Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* 6(9), 1341–1354 (2011).

92 Zhang X, Perica T, Teichmann SA. Evolution of protein structures and interactions from the perspective of residue contact networks. *Curr. Opin. Struct. Biol.* pii: S0959-440X(13)00125-5 (2013).

93 Gao M, Skolnick J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl Acad. Sci. USA* 107(52), 22517–22522 (2010).

94 Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285(5), 2177–2198 (1999).

95 Keskin O, Ma B, Rogale K, Gunasekaran K, Nussinov R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.* 2(2), S24–S35 (2005).

96 Higurashi M, Ishida T, Kinoshita K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.* 17(1), 72–78 (2008).

97 Keskin O, Nussinov R. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure* 15(3), 341–354 (2007).

98 Zhu X, Mitchell JC. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* 79(9), 2671–2683 (2011).

99 Jones S, Thornton JM. Protein-protein interactions: a review of protein dimmer structures. *Prog. Biophys. Mol. Biol.* 63(1), 31–65 (1995).

100 Kleinjung J, Fraternali F. POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Res.* 33(Web Server issue), W342–W346 (2005).

101 Davis F, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21(9), 1901–1907 (2005).

102 Mashiach E, Nussinov R, Wolfson HJ. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Res.* 38 (Web Server issue), W457–W461 (2010).

103 Kim P, Lu L, Xia Y, Gerstein M. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 314(5807), 1938–1941 (2006).

104 Butland G, Peregrín-Alvarez JM, Li J *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature.* 433(7025), 531–537 (2005).

105 Ho H, Milenkovic T, Memisevic V, Aruri J, Przulj N, Ganesan AK. Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC Syst. Biol.* 4(1), 84 (2010).

106 David A, Razali R, Wass MN, Sternberg MJE. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* 33(2), 359–363 (2011).

107 Guo Y, Wei X, Das J *et al.* Dissecting disease inheritance modes in a three-dimensional protein network challenges the guilt-by-association principle. *Am. J. Hum. Genet.* 93(1), 78–89 (2013).

108 Patil A, Kinoshita K, Nakamura H. Hub promiscuity in protein-protein interaction networks. *Int. J. Mol. Sci.* 11(4), 1930–1943 (2010).

109 Tyagi M, Shoemaker BA, Bryant SH, Panchenko AR. Exploring functional roles of multibinding protein interfaces. *Protein Sci.* 18(8), 1674–1683 (2009).

110 Dasgupta B, Nakamura H, Kinjo AR. Distinct roles of overlapping and nonoverlapping regions of hub protein interfaces in recognition of multiple partners. *J. Mol. Biol.* 411(3), 713–727 (2011).

111 Fornili A, Pandini A, Lu H, Fraternali F. Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *J. Chem. Theory Comput.* doi:10.1021/ct400486p (2013) (Epub before print).

112 Zen A, Micheletti C, Keskin O, Nussinov R. Comparing interfacial dynamics in protein-protein complexes: an elastic network approach. *BMC Struct. Biol.* 10, 26 (2010).

113 Spitzer R, Cleves A, Jain A. Surface-based protein binding pocket similarity. *Proteins* 79(9), 2746–2763 (2011).

114 Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.* 7, 469 (2011).

115 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467(7319), 1061–1073 (2010).

116 International HapMap Consortium. The international hapMap project. *Nature* 426(6968), 789–796 (2003).

117 Sherry ST, Ward MH, Kholodov M *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1), 308–311 (2001).

118 Hamosh A. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33(Database issue), D514–D517 (2004).

119 Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39(Database issue), D945–D950 (2011).

120 Stenson PD, Ball EV, Mort M *et al.* Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.* 21(6), 577–581 (2003).

121 Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* 425(21), 3919–3936 (2013).

122 Higueruelo AP, Jubb H, Blundell TL. Protein-protein interactions as druggable targets: recent technological advances. *Curr. Opin. Pharmacol.* 13(5), 791–796 (2013).

Websites

201 Cancer Cell Map. http://cancer.cellmap.org/cellmap

202 Exome Variant Server. http://evs.gs.washington.edu/EVS

Bibliography

[1] Lu, H.-c., Fornili, A. and Fraternali, F. (2013) Protein-protein interaction networks studies and importance of 3D structure knowledge. *Expert Review of Proteomics* **10**(6), 511–520.

[2] Feldman, I., Rzhetsky, A. and Vitkup, D. (2008) Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United* **105**(11), 4323–4328.

[3] Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A. (2010) Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics* **11**, 597.

[4] Carlin, L. M., Evans, R., Milewicz, H., Fernandes, L., Matthews, D. R., Perani, M., Levitt, J., Keppler, M. D., Monypenny, J., Coolen, T., Barber, P. R., Vojnovic, B., Suhling, K., Fraternali, F., Ameer-Beg, S., Parker, P. J., Thomas, N. S. B. and Ng, T. (2011) A targeted siRNA screen identifies regulators of Cdc42 activity at the natural killer cell immunological synapse. *Science Signaling* **4**(201), ra81.

[5] Bakal, C., Linding, R., Llense, F., Heffern, E., Martin-Blanco, E., Pawson, T. and Perrimon, N. (2008) Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science* **322**(5900), 453–456.

[6] Yamada, T. and Bork, P. (2009) Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology* **10**(11), 791–803.

[7] Gursoy, A., Keskin, O. and Nussinov, R. (2008) Topological properties of protein interaction networks from a structural perspective. *Biochemical Society Transactions* **36**(Pt 6), 1398–1403.

[8] Jonsson, P. F. and Bates, P. A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**(18), 2291–2297.

[9] Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. and Ciccarelli, F. D. (2008) Low duplicability and network fragility of cancer genes. *Trends in Genetics* **24**(9), 427–430.

[10] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United* **104**(21), 8685–8690.

[11] Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R. and Kasif, S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United* **101**(9), 2888–2893.

[12] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21 Suppl 1**, i302–i310.

[13] Benfey, P. N. and Mitchell-Olds, T (2008) From Genotype to Phenotype: Systems Biology Meets Natural Variation. *Science* **320**(5875), 495–497.

[14] Arabidopsis Interactome Mapping Consortium, Dreze, M, Carvunis, A. R., Charloteaux, B, Galli, M, Pevzner, S. J., Tasan, M, Ahn, Y. Y., Balumuri, P, Barabási, A.-L., Bautista, V, Braun, P, Byrdsong, D, Chen, H, Chesnut, J. D., Cusick, M. E., Dangl, J. L., Reyes, C de los, Dricot, A, Duarte, M, Ecker, J. R., Fan, C, Gai, L, Gebreab, F, Ghoshal, G, Gilles, P, Gutierrez, B. J., Hao, T, Hill, D. E., Kim, C. J., Kim, R. C., Lurin, C, MacWilliams, A, Matrubutham, U, Milenkovic, T, Mirchandani, J, Monachello, D, Moore, J, Mukhtar, M. S., Olivares, E, Patnaik, S, Poulin, M. M., Przulj, N, Quan, R, Rabello, S, Ramaswamy, G, Reichert, P, Rietman, E. A., Rolland, T, Romero, V, Roth, F. P., Santhanam, B, Schmitz, R. J., Shinn, P, Spooner, W, Stein, J, Swamilingiah, G. M., Tam, S, Vandenhaute, J, Vidal, M, Waaijers, S, Ware, D, Weiner, E. M., Wu, S and Yazaki, J (2011) Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science* **333**(6042), 601–607.

[15] Lee, I., Seo, Y.-S., Coltrane, D., Hwang, S., Oh, T., Marcotte, E. M. and Ronald, P. C. (2011) Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proceedings of the National Academy of Sciences* **108**(45), 18548–18553.

[16] Kuzu, G., Keskin, O., Gursoy, A. and Nussinov, R. (2012) Constructing structural networks of signaling pathways on the proteome scale. *Current Opinion in Structural Biology* **22**(3), 367–377.

[17] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*.

[18] Feng, J., Kim, S.-T., Liu, W., Kim, J. W., Zhang, Z., Zhu, Y., Berens, M., Sun, J. and Xu, J. (2012) An integrated analysis of germline and somatic, genetic and epigenetic alterations at 9p21.3 in glioblastoma. *Cancer* **118**(1), 232–240.

[19] Mathew, J. P., Taylor, B. S., Bader, G. D., Pyarajan, S., Antoniotti, M., Chinnaiyan, A. M., Sander, C., Burakoff, S. J. and Mishra, B. (2007) From bytes to bedside: data integration and computational biology for translational cancer research. *PLOS Computational Biology* **3**(2), e12.

[20] Aloy, P. and Russell, R. B. (2002) Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United* **99**(9), 5896–5901.

[21] Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., Serrano, L. and Russell, R. B. (2004) Structure-based assembly of protein complexes in yeast. *Science* **303**(5666), 2026–2029.

[22] Robinson, C. V., Sali, A. and Baumeister, W. (2007) The molecular sociology of the cell. *Nature* **450**(7172), 973–982.

[23] Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P. and Sali, A. (2007) Determining the architectures of macromolecular assemblies. *Nature* **450**(7170), 683–694.

[24] Mosca, R., Céol, A. and Aloy, P. (2012) Interactome3D: adding structural details to protein networks. *Nature Methods* **10**(1), 47–53.

[25] Meyer, M. J., Das, J., Wang, X. and Yu, H. (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29**(12), 1577–1579.

[26] Bockler, B. and Bateman, A. (2008) Protein interactions in human genetic diseases. *Genome Biology* **9**(1), R9.

[27] Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M. and Yu, H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* **30**(2), 159–164.

[28] Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., Wuchty, S. and Panchenko, A. R. (2013) Cancer Missense Mutations Alter Binding Properties of Proteins and Their Interaction Networks. *PLOS ONE* **8**(6), e66273.

[29] Lees, J. G., Heriche, J. K., Morilla, I, Ranea, J. A. and Orengo, C. A. (2011) Systematic computational prediction of protein interaction networks. *Physical Biology* **8**(3), 035008.

[30] Hooda, Y. and Kim, P. M. (2012) Computational structural analysis of protein interactions and networks. *Proteomics* **12**(10), 1697–1705.

[31] Yue, P., Li, Z. and Moult, J. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology* **353**(2), 459–473.

[32] Studer, R. A., Dessailly, B. H. and Orengo, C. A. (2013) Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical Journal* **449**(3), 581–594.

[33] Satoh, T., Smith, A., Sarde, A., Lu, H.-c., Mian, S., Mian, S., Trouillet, C., Mufti, G., Emile, J.-F., Fraternali, F., Donadieu, J. and Geissmann, F. (2012) B-RAF mutant alleles associated with Langerhans cell histiocytosis, a granulomatous pediatric disease. *PLOS ONE* **7**(4), e33891.

[34] Keskin, O., Ma, B. and Nussinov, R. (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology* **345**(5), 1281–1294.

[35] Bogan, A. A. and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* **280**(1), 1–9.

[36] Liu, J. and Nussinov, R. (2008) Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design. *Proceedings of the National Academy of Sciences* **105**(3), 901–906.

[37] Nussinov, R. and Tsai, C.-J. (2013) Allostery in disease and in drug discovery. *Cell* **153**(2), 293–305.

[38] Shan, Y., Eastwood, M. P., Zhang, X., Kim, E. T., Arkhipov, A., Dror, R. O., Jumper, J., Kuriyan, J. and Shaw, D. E. (2012) Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* **149**(4), 860–870.

[39] Azam, M., Seeliger, M. A., Gray, N. S., Kuriyan, J. and Daley, G. Q. (2008) Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nature Structural & Molecular Biology* **15**(10), 1109–1118.

[40] Blundell, T. L., Sibanda, B. L., Montalvão, R. W., Brewerton, S., Chelliah, V., Worth, C. L., Harmer, N. J., Davies, O. and Burke, D. (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361**(1467), 413–423.

[41] Azzarito, V, Long, K, Murphy, N. S. and Wilson, A. J. (2013) Inhibition of -helix-mediated protein-protein interactions using designed molecules. *Nature Chemistry* **5**, 161–173.

[42] Kastritis, P. L. and Bonvin, A. M. J. J. (2013) On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society Interface* **10**(79), 20120835.

[43] Wells, J. A. and McClendon, C. L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**(7172), 1001–1009.

[44] Winter, C., Henschel, A., Tuukkanen, A. and Schroeder, M. (2012) Protein interactions in 3D: from interface evolution to drug discovery. *Journal of Structural Biology* **179**(3), 347–358.

[45] Keskin, O., Gursoy, A., Ma, B. and Nussinov, R. (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical Reviews* **108**(4), 1225–1244.

[46] Engin, H. B., Gursoy, A., Nussinov, R. and Keskin, O. (2013) Network-Based Strategies Can Help Mono- and Poly-pharmacology Drug Discovery: A Systems Biology View. *Current Pharmaceutical Design*.

[47] Engin, H. B., Keskin, O., Nussinov, R. and Gursoy, A. (2012) A strategy based on protein-protein interface motifs may help in identifying drug off-targets. *Journal of Chemical Information and Modeling* **52**(8), 2273–2286.

[48] Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G. and Nussinov, R. (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics* **138**(3), 333–408.

[49] Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A.-C. and Bork, P. (2013) Systematic identification of proteins that elicit drug side effects. *Molecular Systems Biology* **9**, 663.

[50] Wu, T.-Y., Jen, M.-H., Bottle, A., Molokhia, M., Aylin, P., Bell, D. and Majeed, A. (2010) Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. *Journal of the Royal Society of Medicine* **103**(6), 239–250.

[51] Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. and Vidal, M. (2007) Drug-target network. *Nature Biotechnology* **25**(10), 1119–1126.

[52] Zhang, Q., Petrey, D., Norel, R. and Honig, B. (2010) Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences of the United* **107**(24), 10896–10901.

[53] Kundrotas, P. J., Zhu, Z., Janin, J. and Vakser, I. A. (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences* **109**(24), 9438–9441.

[54] Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A, Rost, B and Sali, A (2001) Reliability of assessment of protein structure prediction methods. *Structure* **10**(3), 435–440.

[55] Schwede, T. (2013) Protein modeling: what happened to the "protein structure gap"? *Structure* **21**(9), 1531–1540.

[56] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw,

R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A. and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**(7084), 637–643.

[57] Stumpf, M. P. H., Thorne, T., Silva, E. de, Stewart, R., An, H. J., Lappe, M. and Wiuf, C. (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United* **105**(19), 6959–6964.

[58] Lehne, B. and Schlitt, T. (2009) Protein-protein interaction databases: keeping up with growing interactomes. *Human Genomics* **3**(3), 291–297.

[59] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V, Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Research* **37**, D767–D772.

[60] Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K. and Tyers, M. (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Research* **41**, D816–D823.

[61] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, D449–D451.

[62] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S. and Hermjakob, H. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Research* **40**, D841–D846.

[63] Ceol, A, Chatr-aryamontri, A, Licata, L, Peluso, D, Briganti, L, Perfetto, L, Castagnoli, L and Cesareni, G (2009) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research* **38**, D532–D539.

[64] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Mering, C. von and Jensen, L. J. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41**, D808–D815.

[65] Braun, P. (2013) Reproducibility restored–on toward the human interactome. *Nature Methods* **10**(4), 301–303.

[66] Varjosalo, M., Sacco, R., Stukalov, A., Drogen, A. van, Planyavsky, M., Hauri, S., Aebersold, R., Bennett, K. L., Colinge, J., Gstaiger, M. and Superti-Furga, G. (2013) Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nature Methods* **10**(4), 307–314.

[67] Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C.,

Vlasblom, J., Dar, V.-u.-N., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M. and Emili, A. (2012) A census of human soluble protein complexes. *Cell* **150**(5), 1068–1081.

[68] Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., Smet, A.-S. de, Venkatesan, K., Rual, J.-F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P. and Vidal, M. (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods* **6**(1), 91–97.

[69] Jain, S. and Bader, G. D. (2010) An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* **11**, 562.

[70] Dazard, J.-E., Saha, S. and Ewing, R. M. (2012) ROCS: a reproducibility index and confidence score for interaction proteomics studies. *BMC Bioinformatics* **13**, 128.

[71] Hakes, L., Pinney, J. W., Robertson, D. L. and Lovell, S. C. (2008) Protein-protein interaction networks and biology—what's the connection? *Nature Biotechnology* **26**(1), 69–72.

[72] Mosca, R., Pons, T., Céol, A., Valencia, A. and Aloy, P. (2013) Towards a detailed atlas of protein-protein interactions. *Current Opinion in Structural Biology*.

[73] Mering, C. von, Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**(6887), 399–403.

[74] Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen,

C., Dumousseau, M., Goll, J., Hancock, R., Hannick, L. I., Jurisica, I., Khadake, J., Lynn, D. J., Mahadevan, U., Perfetto, L., Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stümpflen, V., Tyers, M., Uetz, P., Xenarios, I. and Hermjakob, H. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods* **9**(4), 345–350.

[75] Orchard, S. (2012) Molecular interaction databases. *Proteomics* **12**(10), 1656–1662.

[76] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**(7084), 631–636.

[77] Juan, D. de, Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution. *Nature Reviews Genetics* **14**(4), 249–261.

[78] Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A. and Honig, B. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**(7421), 556–560.

[79] Berman, H. M. (2000) The Protein Data Bank. *Nucleic Acids Research* **28**(1), 235–242.

[80] Finn, R. D., Marshall, M. and Bateman, A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**(3), 410–412.

[81] Stein, A., Céol, A. and Aloy, P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research* **39**(Database issue).

[82] Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S., Bateman, A. and Finn, R. (2012) The Pfam protein families database. *Nucleic Acids Research* **40**(Database issue), D290–D301.

[83] Vogel, C, Teichmann, S and Pereiraleal, J (2005) The Relationship Between Domain Duplication and Recombination. *Journal of Molecular Biology* **346**(1), 355–365.

[84] Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U. and Sali, A. (2007) Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science*, Unit 2.9.

[85] Lewis, T. E., Sillitoe, I., Andreeva, A., Blundell, T. L., Buchan, D. W. A., Chothia, C., Cuff, A., Dana, J. M., Filippis, I., Gough, J., Hunter, S., Jones, D. T., Kelley, L. A., Kleywegt, G. J., Minneci, F., Mitchell, A., Murzin, A. G., Ochoa-Montaño, B., Rackham, O. J. L., Smith, J., Sternberg, M. J. E., Velankar, S., Yeats, C. and Orengo, C. (2013) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Research* **41**(Database issue), D499–D507.

[86] Lees, J., Yeats, C., Redfern, O., Clegg, A. and Orengo, C. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Research* **38**(Database issue), D296–D300.

[87] Shi, J., Blundell, T. L. and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology* **310**(1), 243–257.

[88] Gough, J. and Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research* **30**(1), 268–272.

[89] Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J. E. and Kelley, L. A. (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins* **70**(3), 611–625.

[90] Buchan, D. W. A., Ward, S. M., Lobley, A. E., Nugent, T. C. O., Bryson, K and Jones, D. T. (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Research* **38**(Web Server issue), W563–W568.

[91] Kar, G., Gursoy, A. and Keskin, O. (2009) Human cancer protein-protein interaction network: a structural perspective. *PLOS Computational Biology* **5**(12), e1000601.

[92] Tuncbag, N., Gursoy, A., Nussinov, R. and Keskin, O. (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols* **6**(9), 1341–1354.

[93] Zhang, X., Perica, T. and Teichmann, S. A. (2013) Evolution of protein structures and interactions from the perspective of residue contact networks. *Current Opinion in Structural Biology.*

[94] Gao, M. and Skolnick, J. (2010) Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences of the United* **107**(52), 22517–22522.

[95] Lo Conte, L, Chothia, C and Janin, J (1999) The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology* **285**(5), 2177–2198.

[96] Keskin, O., Ma, B., Rogale, K., Gunasekaran, K and Nussinov, R. (2005) Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Physical Biology* **2**(2), S24–S35.

[97] Krissinel, E. and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**(3), 774–797.

[98] Higurashi, M., Ishida, T. and Kinoshita, K. (2008) Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Science* **17**(1), 72–78.

[99] Keskin, O. and Nussinov, R. (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure* **15**(3), 341–354.

[100] Zhu, X. and Mitchell, J. C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins* **79**(9), 2671–2683.

[101] Jones, S. and Thornton, J. M. (1995) Protein-protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology* **63**(1), 31–65.

[102] Kleinjung, J. and Fraternali, F. (2005) POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Research* **33**(Web Server issue), W342–W346.

[103] Davis, F. and Sali, A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* **21**(9), 1901–1907.

[104] Mashiach, E., Nussinov, R. and Wolfson, H. J. (2010) FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Research* **38**(Web Server issue), W457–W461.

[105] Kim, P., Lu, L., Xia, Y. and Gerstein, M. (2006) Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science* **314**(5807), 1938–1941.

[106] Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**(7025), 531–537.

[107] Ho, H., Milenković, T., Memišević, V., Aruri, J., Pržulj, N. and Ganesan, A. K. (2010) Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC Systems Biology* **4**(1), 84.

[108] David, A., Razali, R., Wass, M. N. and Sternberg, M. J. E. (2011) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation* **33**(2), 359–363.

[109] Guo, Y., Wei, X., Das, J., Grimson, A., Lipkin, S. M., Clark, A. G. and Yu, H. (2013) Dissecting Disease Inheritance Modes in a Three-Dimensional Protein Net-

work Challenges the Guilt-by-Association Principle. *The American Journal of Human Genetics*.

[110] Patil, A., Kinoshita, K. and Nakamura, H. (2010) Hub promiscuity in protein-protein interaction networks. *International Journal of Molecular Sciences* **11**(4), 1930–1943.

[111] Tyagi, M., Shoemaker, B. A., Bryant, S. H. and Panchenko, A. R. (2009) Exploring functional roles of multibinding protein interfaces. *Protein Science* **18**(8), 1674–1683.

[112] Dasgupta, B., Nakamura, H. and Kinjo, A. R. (2011) Distinct roles of overlapping and non-overlapping regions of hub protein interfaces in recognition of multiple partners. *Journal of Molecular Biology* **411**(3), 713–727.

[113] Fornili, A., Pandini, A., Lu, H.-c. and Fraternali, F. (2013) Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *Journal of Chemical Theory and Computation*, 130927133634001.

[114] Zen, A., Micheletti, C., Keskin, O. and Nussinov, R. (2010) Comparing interfacial dynamics in protein-protein complexes: an elastic network approach. *BMC Structural Biology* **10**, 26.

[115] Spitzer, R., Cleves, A. and Jain, A. (2011) Surface-based protein binding pocket similarity. *Proteins* **79**(9), 2746–2763.

[116] Wass, M. N., Fuentes, G., Pons, C., Pazos, F. and Valencia, A. (2011) Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology* **7**, 469.

[117] Nooren, I. M. and Thornton, J. M. (2014) Diversity of protein±protein interactions. *The EMBO Journal*, 1–7.

[118] Janin, J., Miller, S. and Chothia, C. (1988) Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology* **204**(1), 155–164.

[119] Jones, S and Thornton, J. M. (1996) Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United* **93**, 13–20.

[120] Nooren, I. M. A. and Thornton, J. M. (2003) Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology* **325**(5), 991–1018.

[121] The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073.

[122] The International HapMap Consortium (2003) The International HapMap Project. *Nature* **426**(6968), 789–796.

[123] Sherry, S. T., Ward, M. H., Kholodov, M, Baker, J, Phan, L, Smigielski, E. M. and Sirotkin, K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**(1), 308–311.

[124] Hamosh, A (2004) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**(Database issue), D514–D517.

[125] Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. and Futreal, P. A. (2011) COSMIC: mining complete cancer genomes in the

Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**(Database issue), D945–D950.

[126] Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M. and Cooper, D. N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation* **21**(6), 577–581.

[127] Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. and Alexov, E. (2013) Molecular Mechanisms of Disease-Causing Missense Mutations. *Journal of Molecular Biology*.

[128] Higueruelo, A. P., Jubb, H. and Blundell, T. L. (2013) Protein-protein interactions as druggable targets: recent technological advances. *Current Opinion in Pharmacology*.

[129] Psaty, B. M., O'Donnell, C. J., Gudnason, V., Lunetta, K. L., Folsom, A. R., Rotter, J. I., Uitterlinden, A. G., Harris, T. B., Witteman, J. C. M., Boerwinkle, E. and CHARGE Consortium (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation: Cardiovascular Genetics* **2**(1), 73–80.

[130] ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**(5696), 636–640.

[131] Human Genome Structural Variation Working Group, Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., Church, D. M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J. R., Mullikin, J. C., Pritchard, J. K., Sebat, J., Sherry, S. T., Smith, D., Valle, D. and Waterston, R. H. (2007) Completing the map of human genetic variation. *Nature* **447**(7141), 161–165.

[132] Frazer, K. A., Murray, S. S., Schork, N. J. and Topol, E. J. (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**(4), 241–251.

[133] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. and Hurles, M. E. (2006) Global variation in copy number in the human genome. *Nature* **444**(7118), 444–454.

[134] Hehir-Kwa, J., Pfundt, R, Veltman, J. and Leeuw, N de (2013) Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clinical Genetics* **84**(5), 415–421.

[135] Clark, A. G. and Li, J. (2007) Conjuring SNPs to detect associations. *Nature Genetics* **39**(7), 815–816.

[136] Krishnan, V. G. and Ng, P. C. (2012) Predicting cancer drivers: are we there yet? *Genome Medicine* **4**(11), 88.

[137] Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. and Hay, S. I. (2005) The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature* **434**(7030), 214–217.

[138] Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J. and Hay, S. I. (2010) Global distribution of the sickle cell gene

and geographical confirmation of the malaria hypothesis. *Nature Communications* **1**, 104.

[139] Kwiatkowski, D. P. (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics* **77**(2), 171–192.

[140] Hedrick, P (2004) Estimation of relative fitnesses from relative risk data and the predicted future of haemoglobin alleles S and C. *Journal of Evolutionary Biology* **17**(1), 221–224.

[141] Aneni, E. C., Hamer, D. H. and Gill, C. J. (2013) Systematic review of current and emerging strategies for reducing morbidity from malaria in sickle cell disease. *Tropical Medicine & International Health* **18**(3), 313–327.

[142] Mitchell-Olds, T., Willis, J. H. and Goldstein, D. B. (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics* **8**(11), 845–856.

[143] Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., Watson, M. S., Martin, C. L. and Ledbetter, D. H. (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *The American Journal of Human Genetics* **86**(5), 749–764.

[144] Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. and Chia, K. S. (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* **55**(7), 403–415.

[145] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. and Hirschhorn, J. N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**(5), 356–369.

[146] Ng, P. C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**(13), 3812–3814.

[147] Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. and Chan, A. P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PloS ONE* **7**(10), e46688.

[148] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nature Methods* **7**(4), 248–249.

[149] Thomas, P. D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United* **101**(43), 15398–15403.

[150] Bromberg, Y. and Rost, B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research* **35**(11), 3823–3835.

[151] Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* **33**(Web Server issue), W306–10.

[152] Al-Numair, N. S. and Martin, A. C. R. (2013) The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* **14 Suppl 3**, S4.

[153] Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M. and Gaunt, T. R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation* **34**(1), 57–65.

[154] Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**(17), e118.

[155] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K. and Ding, L. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Research* **22**(8), 1589–1598.

[156] Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B. and Karchin, R. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69**(16), 6660–6667.

[157] Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z. and Cavet, G. (2010) Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation* **31**(3), 264–271.

[158] Kaminker, J. S., Zhang, Y., Watanabe, C. and Zhang, Z. (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* **35**(Web Server issue), W595–8.

[159] Marshall, C. R. and Scherer, S. W. (2012) Detection and characterization of copy number variation in autism spectrum disorder. *Methods in Molecular Biology* **838**, 115–135.

[160] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., Grassi, A. de, Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E. and Dermitzakis, E. T. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**(5813), 848–853.

[161] Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O. and Snyder, M. (2010) Variation in transcription factor binding among humans. *Science* **328**(5975), 232–235.

[162] Harewood, L., Chaignat, E. and Reymond, A. (2012) Structural variation and its effect on expression. *Methods in Molecular Biology* **838**, 173–186.

[163] Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A. and Eichler, E. E. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *The American Journal of Human Genetics* **84**(2), 148–161.

[164] Nguyen, H.-H., Hannemann, F., Hartmann, M. F., Malunowicz, E. M., Wudy, S. A. and Bernhardt, R. (2010) Five novel mutations in CYP11B2 gene detected in patients with aldosterone synthase deficiency type I: Functional characterization and structural analyses. *Molecular Genetics and Metabolism* **100**(4), 357–364.

[165] Murray, M. M., Krone, M. G., Bernstein, S. L., Baumketner, A., Condron, M. M., Lazo, N. D., Teplow, D. B., Wyttenbach, T., Shea, J.-E. and Bowers, M. T. (2009) Amyloid beta-protein: experiment and theory on the 21-30 fragment. *The Journal of Physical Chemistry B* **113**(17), 6041–6046.

[166] Dunker, A. K., Obradovic, Z, Romero, P and Garner, E. C. (2000) Intrinsic protein disorder in complete genomes. *Genome Informatics* **11**, 161–171.

[167] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology* **337**(3), 635–645.

[168] Gsponer, J. and Madan Babu, M (2009) The rules of disorder or why disorder rules. *Progress in Biophysics and Molecular Biology* **99**(2-3), 94–103.

[169] Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P, Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J, Nissen, M. S., Reeves, R, Kang, C, Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W, Garner, E. C. and Obradovic, Z (2001) Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling* **19**(1), 26–59.

[170] Dunker, A. K. and Obradovic, Z. (2001) The protein trinity–linking function and disorder. *Nature Biotechnology* **19**(9), 805–806.

[171] Uversky, V. N. (2002) Natively unfolded proteins: A point where biology waits for physics. *Protein Science* **11**(4), 739–756.

[172] Huth, J. R., Bewley, C. A., Nissen, M. S., Evans, J. N., Reeves, R, Gronenborn, A. M. and Clore, G. M. (1997) The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. *Nature Structural Biology* **4**(8), 657–665.

[173] Hibbert, R. G., Teriete, P., Grundy, G. J., Beavil, R. L., Reljic, R., Holers, V. M., Hannan, J. P., Sutton, B. J., Gould, H. J. and McDonnell, J. M. (2005) The structure of human CD23 and its interactions with IgE and CD21. *The Journal of Experimental Medicine* **202**(6), 751–760.

[174] Longhi, S., Receveur-Bréchot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S. and Canard, B. (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *The Journal of Biological Chemistry* **278**(20), 18638–18648.

[175] Disfani, F. M., Hsu, W.-L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., Uversky, V. N. and Kurgan, L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* **28**(12), i75–83.

[176] Ward, J., McGuffin, L., Bryson, K., Buxton, B. and Jones, D. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**(13), 2138–2139.

[177] Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. and Uversky, V. N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta* **1804**(4), 996–1010.

[178] Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I. and Sussman, J. L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**(16), 3435–3438.

[179] Ferron, F., Longhi, S., Canard, B. and Karlin, D. (2006) A practical overview of protein disorder prediction methods. *Proteins* **65**(1), 1–14.

[180] He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. and Dunker, K. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Research* **19**(8), 929–949.

[181] Dinkel, H, Michael, S, Weatheritt, R. J., Davey, N. E., Van Roey, K, Altenberg, B, Toedt, G, Uyar, B, Seiler, M, Budd, A, Jodicke, L, Dammert, M. A., Schroeter, C, Hammer, M, Schmidt, T, Jehl, P, McGuigan, C, Dymecka, M, Chica, C, Luck, K, Via, A, Chatr-aryamontri, A, Haslam, N, Grebnev, G, Edwards, R. J., Steinmetz, M. O., Meiselbach, H, Diella, F and Gibson, T. J. (2011) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Research* **40**(D1), D242–D251.

[182] Romero, P, Obradovic, Z, Li, X, Garner, E. C., Brown, C. J. and Dunker, A. K. (2001) Sequence complexity of disordered protein. *Proteins* **42**(1), 38–48.

[183] Das, R. K. and Pappu, R. V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **110**(33), 13392–13397.

[184] Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z. and Dunker, A. K. (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Research* **35**(Database issue), D786–93.

[185] Vucetic, S., Brown, C., Dunker, K. and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins* **52**(4), 573–584.

[186] Cumberworth, A., Lamour, G., Babu, M. M. and Gsponer, J. (2013) Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal* **454**(3), 361–369.

[187] Winter, J, Ilbert, M, Graf, P. C. F., Ozcelik, D and Jakob, U (2008) Bleach activates a redox-regulated chaperone by oxidative protein unfolding. *Cell* **135**(4), 691–701.

[188] The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073.

[189] The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661–678.

[190] Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**(6), 695–701.

[191] Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A, Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K.-S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. and Wang, J. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**(7218), 60–65.

[192] Plunkett, J., Doniger, S., Morgan, T., Haataja, R., Hallman, M., Puttonen, H., Menon, R., Kuczynski, E., Norwitz, E., Snegovskikh, V., Palotie, A., Peltonen, L., Fellman, V., DeFranco, E. A., Chaudhari, B. P., Oates, J., Boutaud, O., McGre-

gor, T. L., McElroy, J. J., Teramo, K., Borecki, I., Fay, J. C. and Muglia, L. J. (2010) Primate-specific evolution of noncoding element insertion into PLA2G4C and human preterm birth. *BMC Medical Genomics* **3**, 62.

[193] Chen, Y., Zhu, J., Lum, P. Y., Yang, X., Pinto, S., MacNeil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., Leonardson, A., Castellini, L. W., Wang, S., Champy, M.-F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T. A., Lusis, A. J. and Schadt, E. E. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**(7186), 429–435.

[194] Gnad, F., Baucom, A., Mukhyala, K., Manning, G. and Zhang, Z. (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14 Suppl 3**, S7.

[195] Bhardwaj, N., Abyzov, A., Clarke, D., Shou, C. and Gerstein, M. B. (2011) Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Science* **20**(10), 1745–1754.

[196] Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K. and Uversky, V. N. (2006) Analysis of Molecular Recognition Features (MoRFs). *Journal of Molecular Biology* **362**(5), 1043–1059.

[197] Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology* **323**(3), 573–584.

[198] Vacic, V. and Iakoucheva, L. M. (2011) Disease mutations in disordered regions—exception to the rule? *Molecular BioSystems* **8**(1), 27–32.

[199] Pajkos, M., Mészáros, B., Simon, I. and Dosztányi, Z. (2012) Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Molecular BioSystems* **8**(1), 296–307.

[200] Joosten, R. P., Beek, T. A. H. te, Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C. and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**(Database issue), D411–D419.

[201] Leu, S, Felten, S von, Frank, S, Vassella, E, Vajtai, I, Taylor, E, Schulz, M, Hutter, G, Hench, J, Schucht, P, Boulay, J. L. and Mariani, L (2013) IDH/MGMT-driven molecular classification of low-grade glioma is a strong predictor for long-term survival. *Neuro-Oncology* **15**(4), 469–479.

[202] Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K. N., Shinjo, S. M. O., Yan, H., Riggins, G. J., Bigner, D. D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. and Kinzler, K. W. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**(5897), 1807–1812.

[203] Magrane, M. and Uniprot Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, bar009.

[204] Minguez, P., Letunic, I., Parca, L. and Bork, P. (2013) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Research* **41**(Database issue), D306–D311.

[205] Wan, P. T. C., Garnett, M. J., Roe, S. M., Lee, S., Niculescu-Duvaz, D., Good, V. M., Jones, C. M., Marshall, C. J., Springer, C. J., Barford, D., Marais, R. and Cancer Genome Project (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**(6), 855–867.

[206] Gough, J, Karplus, K, Hughey, R and Chothia, C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology* **313**(4), 903–919.

[207] Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009) The cancer genome. *Nature* **458**(7239), 719–724.

[208] Altschul, S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389–3402.

[209] Petersen, M., Andersen, J. T., Jimenez-Solem, E., Broedbaek, K., Hjelvang, B. R., Henriksen, T., Frandsen, E., Forman, J. L., Torp-Pedersen, C., Køber, L. and Poulsen, H. E. (2012) Effect of the Arg389Gly β 1-adrenoceptor polymorphism on plasma renin activity and heart rate, and the genotype-dependent response to metoprolol treatment. *Clinical and Experimental Pharmacology and Physiology* **39**(9), 779–785.

[210] Gray, K. A., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W. and Bruford, E. A. (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Research* **41**(Database issue), D545–52.

[211] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2008. URL: `http://www.R-project.org`.

[212] Yu, H., Jansen, R., Stolovitzky, G. and Gerstein, M. (2007) Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics* **23**(16), 2163–2173.

[213] Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., Gardner, M., Laiho, K., Legge, D., Magrane, M., Pichler, K., Poggioli, D., Sehra, H., Auchincloss, A., Axelsen, K., Blatter, M.-C., Boutet, E., Braconi-Quintaje, S., Breuza, L., Bridge, A., Coudert, E., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., James, J., Jimenez, S., Jungo, F., Keller, G., Lemercier, P., Lieberherr, D., Masson, P., Moinat, M., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Bougueleret, L., Argoud-Puy, G., Cusin, I., Duek-Roggli, P., Xenarios, I. and Apweiler, R. (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Research* **40**(Database issue), D565–D570.

[214] Vogel, C., Berzuini, C., Bashton, M., Gough, J. and Teichmann, S. A. (2004) Supra-domains: evolutionary units larger than single protein domains. *Journal of Molecular Biology* **336**(3), 809–823.

[215] Vogel, C. and Chothia, C. (2006) Protein family expansions and biological complexity. *PLOS Computational Biology* **2**(5), e48.

[216] Mosca, R., Céol, A., Stein, A., Olivella, R. and Aloy, P. (2013) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*.

[217] Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M. and

Orengo, C. A. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research* **41**(Database issue), D490–8.

[218] Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research* **36**(Database issue), D419–25.

[219] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Research* **38**(Database issue), D211–D222.

[220] Finn, R. D., Clements, J. and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**(Web Server issue), W29–W37.

[221] Zhang, Z, Schwartz, S, Wagner, L and Miller, W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**(1-2), 203–214.

[222] Notredame, C, Higgins, D. G. and Heringa, J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**(1), 205–217.

[223] Eswar, N., Eramian, D., Webb, B., Shen, M.-Y. and Sali, A. (2008) Protein structure modeling with MODELLER. *Methods in Molecular Biology* **426**, 145–159.

[224] Scharner, J., Lu, H.-c., Fraternali, F., Ellis, J. A. and Zammit, P. S. (2013) Mapping disease-related missense mutations in the immunoglobulin-like fold domain of

lamin A/C reveals novel genotype-phenotype associations for laminopathies. *Proteins*.

[225] Cochran, A. G. (2000) Antagonists of protein–protein interactions. *Chemistry & Biology*.

[226] Hopkins, A. L. and Groom, C. R. (2002) The druggable genome. *Nature Reviews. Drug Discovery* **1**(9), 727–730.

[227] Wolfe, R. and Hanley, J. (2002) If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal*.

[228] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. R. (2004) A census of human cancer genes. *Nature Reviews Cancer* **4**(3), 177–183.